IJCSIS Vol. 8 No. 8, November 2010 ISSN 1947-5500

International Journal of Computer Science & Information Security

© IJCSIS PUBLICATION 2010

Editorial Message from Managing Editor

The International Journal of Computer Science and Information Security (IJCSIS) is a monthly periodical on research articles in general computer science and information security which provides a distinctive technical perspective on novel technical research work, whether theoretical, applicable, or related to implementation. This journal is supported by a dynamic team of high-calibre editorial members and reviewers worldwide.

Target Audience: IT academics, university IT faculties; and business people concerned with computer science and security; industry IT departments; government departments; the financial industry; the mobile industry and the computing industry.

Coverage includes: security infrastructures, network security: Internet security, content protection, cryptography, steganography and formal methods in information security; multimedia systems, software, information systems, intelligent systems, web services, data mining, wireless communication, networking and technologies, innovation technology and management.

Thanks for your contributions in November 2010 issue and we are grateful to the reviewers for providing valuable comments. IJCSIS November 2010 Issue (Vol. 8, No. 8) has paper acceptance rate of up to 30%.

Available at http://sites.google.com/site/ijcsis/

IJCSIS Vol. 8, No. 8, November 2010 Edition ISSN 1947-5500 © IJCSIS, USA.

Abstracts Indexed by (among others):



LICSIS EDITORIAL BOARD

Dr. Gregorio Martinez Perez

Associate Professor - Professor Titular de Universidad, University of Murcia (UMU), Spain

Dr. M. Emre Celebi,

Assistant Professor, Department of Computer Science, Louisiana State University in Shreveport, USA

Dr. Yong Li

School of Electronic and Information Engineering, Beijing Jiaotong University, P. R. China

Prof. Hamid Reza Naji

Department of Computer Enigneering, Shahid Beheshti University, Tehran, Iran

Dr. Sanjay Jasola

Professor and Dean, School of Information and Communication Technology, Gautam Buddha University

Dr Riktesh Srivastava

Assistant Professor, Information Systems, Skyline University College, University City of Sharjah, Sharjah, PO 1797, UAE

Dr. Siddhivinayak Kulkarni

University of Ballarat, Ballarat, Victoria, Australia

Professor (Dr) Mokhtar Beldjehem

Sainte-Anne University, Halifax, NS, Canada

Dr. Alex Pappachen James, (Research Fellow)

Queensland Micro-nanotechnology center, Griffith University, Australia

Dr. T.C. Manjunath,

ATRIA Institute of Tech, India.

TABLE OF CONTENTS

1. Paper 12101008: A Brief Survey on RFID Security and Privacy Issues (pp. 1-10)

Mohammad Tauhidul Islam

Department of Mathematics and Computer Science, University of Lethbridge, Alberta, Canada T1K 3M4.

2. Paper 26101031: An Improved Fuzzy Time Series Model For Forecasting (pp. 11-19)

Ashraf K. Abd-Elaal, Department of Computer and Information Sciences, The High Institute of Computer Science, Sohag, Egypt

Hesham A. Hefny, Department of Computer and Information Sciences, Institute of Statistical Studies and Research, Cairo University, Egypt

Ashraf H. Abd-Elwahab, Department of Computer Sciences, Electronics Research Institute National Center for Research, Cairo, Egypt

3. Paper 30101046: The 2D Image-Based Anthropologic Measurement By Using Chinese Medical Acupuncture And Human Body Slice Model (pp. 20-29)

Sheng-Fuu Lin, Institute of Electrical Control Engineering, National Chiao Tung University, 1001 Ta Hsueh Road, Hsinchu City, Taiwan 300, ROC

Shih-Che Chien, Institute of Electrical Control Engineering, National Chiao Tung University, 1001 Ta Hsueh Road, Hsinchu City, Taiwan 300, ROC

Kuo-Yu Chiu, Institute of Electrical Control Engineering, National Chiao Tung University, 1001 Ta Hsueh Road, Hsinchu City, Taiwan 300, ROC

4. Paper 31101073: A Fast Fractal Image Encoding Based On Haar Wavelet Transform (pp. 30-36)

Sofia Douda, Département de Mathématiques et Informatique & ENIC, Faculté des Sciences et Techniques, Université Hassan 1^{er}, Settat, Morocco.

Abdallah Bagri, ENIC, Faculté des Sciences et Techniques, Université Hassan 1^{er}, Settat, Morocco. Abdelhakim El Imrani, LCS, Faculté des Sciences, Université Mohammed V, Rabat, Morocco

5. Paper 31101082: A New Noise Estimation Technique of Speech Signal by Degree of Noise Refinement (pp. 37-43)

Md. Ekramul Hamid, College of Computer Science, King Khalid University, Abha, Kingdom of Saudi Arabia

Md. Zashim Uddin, Department of Computer Science and Engg., University of Rajshahi, Rajshahi, Bangladesh.

Md. Humayun Kabir Biswas, College of Computer Science, King Khalid University, Abha, Kingdom of Saudi Arabia

Somlal Das, Dept. of Computer Science, University of Rajshahi, Rajshahi, Bangladesh

6. Paper 31101060: Scalable Video Coding in Online Video transmission with Bandwidth Limitation (pp. 44-47)

Sima Ahmadpour, Salah Noori Saleh, Omar Amer Abouabdalla, Mahmoud Baklizi, Nibras Abdullah National Advanced IPv6 Center of Excellence, University Science Malaysia, Penang, Malaysia

7. Paper 30101053: Off-Line Handwritten Signature Retrieval using Curvelet Transforms (pp. 48-51)

M. S. Shirdhonkar, Dept. of Computer Science and Engineering, B.L.D.E.A's College of Engineering and Technology, Bijapur, India

Manesh Kokare, Dept. of Electronics and Telecommunication, S.G.G.S Institute of Engineering and Technology, Nanded, India

8. Paper 14101014: Low Complexity MMSE Based Channel Estimation Technique for LTE OFDMA Systems (pp. 52-56)

Md. Masud Rana, Department of Electronics and Radio Engineering Kyung Hee University, South Korea Abbas Z. Kouzani, School of Engineering, Deakin University, Geelong, Victoria 3217, Australia

9. Paper 27101033: Survey: RTCP Feedback In A Large Streaming Sessions (pp. 57-62)

Adel Nadhem Naeem , Ali Abdulqader Bin Salem , Mohammed Faiz Aboalmaaly , and Sureswaran Ramadass

National Advanced IPv6 Centre, Universiti Sains Malaysia, Pinang, Malaysia

10. Paper 27101034: Performance Analysis Of Nonlinear Distortions For Downlink MC-CDMA Systems (pp. 63-70)

Labib Francis Gergis

Misr Academy for Engineering and Technology, Mansoura, Egypt

11. Paper 16101016: Channel Estimation Algorithms, Complexities and LTE Implementation Challenges (pp. 71-76)

Md. Masud Rana

Department of Electronics and Communication Engineering, Khulna University of Engineering and Technology, Khunla, Bangladesh

12. Paper 18101017: Implementation Of Wavelet And RBF For Power Quality Disturbance Classification (pp. 77-82)

Pramila P^{1} , Puttamadappa C^{2} and S. Purushothaman ³

13. Paper 28091037: GA-ANN based Dominant Gene Prediction in Microarray Dataset (pp. 83-91)

Manaswini Pradhan, Lecturer, P.G. Department of Information and Communication Technology, Fakir Mohan University, Orissa, India

Dr. Sabyasachi Pattnaik, Reader, P.G. Department of Information and Communication Technology, Fakir Mohan University, Orissa, India.

Dr. B. Mittra, Reader, School of Biotechnology, Fakir Mohan University, Orissa, India

Dr. Ranjit Kumar Sahu, Assistant Surgeon, Post Doctoral Department of Plastic and Reconstructive Surgery, S.C.B. Medical College, Cuttack, Orissa, India

¹ Department of Electrical & Electronics Engineering , Bangalore Institute Of Technology , Bangalore, India

² Department of Electronics & Communication Engineering, SJB Institute Of Technology, Bangalore, India

³ Sun College Of Engineering & Technology, Sunnagar, Kanyakumari, Tamilnadu, India

14. Paper 29091039: Therapeutic Diet Prediction for Integrated Mining of Anemia Human Subjects using Statistical Techniques (pp. 92-95)

Sanjay Choudhary, Department of Mathematics & Computer Science, Govt. Narmada P.G. Mahavidyalaya Hoshangabad, India

Abha Wadhwa, Department of Computer Science & Application, Govt Girls P.G. College, Hoshangabad, India

Kamal Wadhwa, Department of Mathematics & Computer Science, Govt. Narmada P.G. Mahavidyalaya Hoshangabad, India

Anjana Mishra, Department of Mathematics & Computer Science, Govt. Narmada P.G. Mahavidyalaya Hoshangabad, India

15. Paper 29101041: Improve the Test Case Design of Object Oriented Software by Refactoring (pp. 96-100)

Divya Prakash Shrivastava , Department of Computer Science, Al Jabal Al Garbi University, Zawya, LIBYA

R.C. Jain, Department of Computer Application, Samrat Ashoka Technological Institute, Vidisha, INDIA

16. Paper 29101042: Extraction of Information from Images using Dewrapping Techniques (pp. 101-109)

Khalid Nazim S. A., Research Scholar, Singhania University, Rajasthan, India. Dr. M.B. Sanjay Pande, Professor and Head, Department of Computer Science & Engineering, VVIET, Mysore, India

17. Paper 29101043: Secured Authentication Protocol System Using Images (pp. 110-116)

G. Arumugam, Prof. & Head, Computer Science Department, Madurai Kamaraj University, Madurai, India

R. Sujatha, Research Associate, SSE Project, Department of Computer Science, Madurai Kamaraj University, Madurai, India

18. Paper 29101044: SIP and RSW: A Comparative Evaluation Study (pp. 117-119)

Mahmoud Baklizi, Nibras Abdullah, Omar Abouabdalla, Sima Ahmadpour National Advanced IPv6 Centre of Excellence, Universiti Sains Malaysia, Penang, Malaysia

19. Paper 30091053: A role oriented requirements analysis for ERP implementation in health care Organizations (pp. 120-124)

Kirti Pancholi, Acropolis Institute of Pharmaceutical Education and Research, Indore, MP, India Durgesh Kumar Mishra, Acropolis Institute of Technology and Research, Indore, MP, India

20. Paper 30101050: Fuzzy expert system for evaluation of students and online exams (pp. 125-130)

Mohammed E. Abd-Alazeem, Computer science department, Faculty of computers and information, Mansoura, Egypt.

Sherief I. Barakat, Information system department, Faculty of computers and information, Mansoura , Egypt.

21. Paper 30101051: Intelligent Controller for Networked DC Motor Control (pp. 131-137)

B. Sharmila, Department of EIE, Sri Ramakrishna Engineering College, Coimbatore, India N. Devarajan, Department of EEE, Government College of Tech, Coimbatore, India

22. Paper 31101062: A Novel LTCC Bandpass Filter for UWB Applications (pp. 138-140)

Thirumalaivasan K. and Nakkeeran R.

Department of Electronics and Communication Engineering, Pondicherry Engineering College, Puducherry-605014, India

23. Paper 25101027: Retrieval of Bitmap Compression History (pp. 141-146)

Salma Hamdy, Haytham El-Messiry, Mohamed Roushdy, Essam Kahlifa Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt

24. Paper 11101007: Steganography and Error-Correcting Codes (pp. 147-149)

M.B. Ould MEDENI and El Mamoun SOUIDI

Laboratory of Mathematic Informatics and Applications, University Mohammed V-Agdal, Faculty of Sciences, Rabat, BP 1014, Morocco

25. Paper 29091047: A Comparative Study on Kakkot Sort and Other Sorting Methods (pp. 150-155)

Rajesh Ramachandran, HOD, Department of Computer Science, Naipunnya Institute of Management & Information Technology, Pongam, Kerala Dr. E. Kirubakaran, Sr. DGM(Outsourcing), BHEL, Trichy

26. Paper 27101037: A Generalization of the PVD Steganographic Method (pp. 156-159)

M. B. Ould MEDENI, Laboratory of Mathematic Informatics and Applications, University Mohammed V-Agdal, Faculty of Sciences, Rabat, BP 1014, Morocco

El Mamoun SOUIDI, Laboratory of Mathematic Informatics and Applications, University Mohammed V-Agdal, Faculty of Sciences, Rabat , BP 1014, Morocco

27. Paper 12101009: Implementation of Polynimial Neural Network in Web Usage Mining (pp. 160-167)

S. Santhi, Research Scholar, Mother Teresa Women's University, Kodaikanal, India Dr. S. Purushothaman, Principal, Sun college of Engineering and Technology, Nagarkoil, India

28. Paper 14101013: Efficient Probabilistic Classification Methods for NIDS (pp. 168-172)

S.M. Aqil Burney, Meritorious Professor Department of Computer Science, University of Karachi, Karachi-Pakistan

M. Sadiq Ali Khan, Assistant Professor Department of Computer Science, University of Karachi, Karachi-Pakistan.

Jawed Naseem, Principal Scientific Officer-PARC

29. Paper 16101015: A Survey on Digital Image Enhancement Techniques (pp. 173-178)

V. Saradhadevi, Research scholar, Karpagam University, Coimbatore, India. Dr. V. Sundaram, Director of MCA, karpagam Engineering College, Coimbatore, India.

30. Paper 18101019: A Survey on Designing Metrics suite to Asses the Quality of Ontology (pp. 179-184)

K. R. Uthayan, Department of Information Technology, SSN College of Engineering, Chennai, India G. S. Anandha Mala, Professor & Head, Department of Computer Science & Engineering, St. Joseph's College of Engineering, Chennai, India

31. Paper 20101021: An Anomaly-Based Network Intrusion Detection System Using Fuzzy Logic (pp. 185-193)

R. Shanmugavadivu, Assistant professor, Department of Computer Science, PSG College of Arts & Science, Coimbatore.

Dr. N. Nagarajan, Principal, Coimbatore Institute of Engineering and Information Technology, Coimbatore.

32. Paper 25101030: Blemish Tolerance in Cellular Automata And Evaluation Reliability (pp. 194-200)

Rogheye parikhani, Engineering Department, Islamic Azad University, Tabriz branch, Tabriz, Iran Mohmad teshnelab, Department of Controls Engineering, Faculty of Electrical and Computer Engineering, KN Toosi University of Technology, Tehran, Iran Shahram babaye, Engineering Department, Islamic Azad University, Tabriz branch, Tabriz, Iran

33. Paper 30101048: Feed Forward Neural Network Algorithm for Frequent Patterns Mining (pp. 201-205)

Amit Bhagat, Department of Computer Applications
Dr. Sanjay Sharma, Associate Prof. Deptt. of Computer Applications
Dr. K.R.Pardasani, Professor Deptt. of Mathematics

Maulana Azad National Institute of Technology, Bhopal (M.P.)462051, India

34. Paper 30101057: An Efficient Vector Quantization Method for Image Compression with Codebook generation using Modified K-Means (pp. 206-212)

S. Sathappan, Associate Professor of Computer Science, Erode Arts and Science College, Erode-638 009. Tamil Nadu. India.

35. Paper 31101064: Optimization of work flow execution in ETL using Secure Genetic Algorithm (pp. 213-222)

Raman Kumar, Saumya Singla, Sagar Bhalla and Harshit Arora Department of Computer Science and Engineering, D A V Institute of Engineering and Technology, Jalandhar, Punjab, India.

36. Paper 31101066: A Survey of: 3D Protein Structure Comparison and Retrieval Methods (pp. 223-227)

Muhannad A. Abu-Hashem, Nur'Aini Abdul Rashid, Rosni Abdullah, Hesham A. Bahamish School of Computer Science, Universiti Sains Malaysia USM, Penang, Malaysia

37. Paper 31101084: The Impact of Speed on the Performance of Dynamic Source Routing in Mobile Ad-Hoc Networks (pp. 228-233)

Naseer Ali Husieen, Osman B Ghazali, Suhaidi Hassan, Mohammed M. Kadhum Internetworks Research Group, College of Arts and Sciences, University Utara Malaysia, 06010 UUM Sintok, Malaysia

38. Paper xxxxx: Multidimensionality in Agile Software Development (pp. 234-238)

Ashima, Assistant Professor, Computer Science and Engineering Department, Thapar University, Patiala Dr. Himanshu Aggarwal, Associate Professor. Faculty of Computer Engineering, Punjabi University, Patiala.

39. Paper 31101093: Aggregating Intrusion Detection System Alerts Based on *Row Echelon Form* Concept (pp. 239-242)

Homam El-Taj, Omar Abouabdalla, Ahmed Manasrah, Moein Mayeh, Mohammed Elhalabi National Advanced IPv6 Center (NAv6) UNIVERSITI SAINS MALAYSIA Penang, Malaysia 11800

40. Paper 31101056: Evaluation of Vision based Surface Roughness using Wavelet Transforms with Neural Network Approach (pp. 243-252)

T.K. Thivakaran Research scholar, MS University, Thirunelveli Dr. RM. Chandrasekaran, Professor, Annamalai University, Chidambaram

41. Paper 31101059: An In-Depth Study on Requirement Engineering (pp. 253-262)

Mohammad Shabbir Hasan, Abdullah Al Mahmood, Farin Rahman, Sk. Md. Nahid Hasan, Panacea Research Lab, Dhaka, Bangladesh

42. Paper 31101072: GCC license plates detection and recognition using morphological filtering and neural networks (pp. 263-269)

Mohammed Deriche

Electrical Engineering Department, King Fahd University of Petroleum and Minerals, Dhahran, 31261 Saudi Arabia.

43. Paper 11101003: Combined Algorithm of Particle Swarm Optimization (pp. 270-276)

Narinder Singh *, S.B. Singh*, and J.C.Bansal**

- *Department of Mathematics, Punjabi University, Patiala, Punjab, INDIA
- ** ABV-Indian Institute of Information Technology and Management-Gwalior (M.P),-INDIA

44. Paper 31101081: Optimal Solution of 2-Dim Rectangle Packing Problem based on Particle Swarm Optimization (pp. 277-283)

Narinder Singh *, S.B. Singh*, and J.C.Bansal**

- *Department of Mathematics, Punjabi University, Patiala, Punjab, INDIA
- ** ABV-Indian Institute of Information Technology and Management-Gwalior (M.P),-INDIA

45. Paper 29101040: Localization Accuracy Improved Methods Based on Adaptive Weighted Centroid Localization Algorithm in Wireless Sensor Networks (pp. 284-288)

Chang-Woo Song, Jun-Ling Ma, Jung-Hyun Lee, Department of Information Engineering, INHA University, Incheon, Korea.

Kyung-Yong Chung, Department of Computer Information Engineering, Sangji University, Wonju, Korea Kee-Wook Rim, Department of Computer and Information Science, Sunmoon University, Asan, Korea

46. Paper 31101068: A Novel Hybridization of ABC with CBR for Pseudoknotted RNA Structure (pp. 289-299)

Ra'ed M. Al-Khatib, Nur'Aini Abdul Rashid and Rosni Abdullah School of Computer Science, Universiti Sains Malaysia USM, Penang, Malaysia

47. Paper 31101070: Hybrid JPEG Compression Using Histogram Based Segmentation (pp. 300-306)

M.Mohamed Sathik, Department of Computer Science, Sadakathullah Appa College, Tirunelveli, India. K.Senthamarai Kannan, Department of Statistics, Manonmaniam Sundaranar University, Tirunelveli, India. Y.Jacob Vetha Raj, Department of Statistics, Manonmaniam Sundaranar University, Tirunelveli, India.

A Brief Survey on RFID Security and Privacy Issues

Mohammad Tauhidul Islam

Department of Mathematics and Computer Science University of Lethbridge, Alberta, Canada T1K 3M4.

Abstract-Radio Frequency IDentification (RFID) security and privacy are exciting research areas that involve affluent interactions among many disciplines like signal processing, supply-chain logistics, hardware design, privacy rights and cryptography. There remain connections to be explored between the work surveyed here and other areas of study. This paper explores by highlighting a few of these. The majority of the articles treated in this survey explore security and privacy as an issue between RFID tags and readers and also compare with other technologies such as Barcode. Of course, tags and readers lie at the periphery of a full-scale RFID system. Many of the attendant data-security problems like that of authenticating readers to servers involve already familiar data-security protocols. This paper also mentions key management, costing, tag collision for RFID and identifies PIN distribution for tags as one such potential problem.

Keywords-RFID; Privacy and security; RFID tags; RFID readers

I. INTRODUCTION

RFID technology uses radio-frequency waves to automatically identify people or objects. There are several methods of identification, but the most common is to store a serial number that identifies a person or object, and perhaps other information, on a microchip that is attached to an antenna (the chip and the antenna together are called an RFID transponder or an RFID tag). The antenna enables the chip to transmit the identification information to a reader. The reader converts the radio waves reflected back from the RFID tag into digital information that can then be passed on to computers that can make use of it. RFID is automatic and fast and will replace the barcode system in the near future. The big difference between RFID and barcodes is line-of-sight technology. That is, a scanner has to see the barcode to read it, which means people usually have to adjust the barcode toward a scanner for it to be read, RFID by contrast, does not require line of sight. RFID tags can be read as long as they are within range of a reader. RFID is a proven technology that has been around since at least the 1970s. Up until now, it has been too expensive and too limited to be realistic for commercial applications. But if the cost associated with making tags is reduced enough, they can solve many of the problems associated with barcodes and bring much more benefit. Both the size and cost of RFID tags have been continuously decreasing. With potentially significant applications and the cheap price of RFID technology, it is predictable that every moving object could be tagged in the near future. In recent years, since Wal-Mart originated and applied RFID technology in supply chain management, RFID has been widely used in many different fields such as defence and military, postal package tracking, aviation industry, health care and baggage and passenger tracing in airport etc. The use of an RFID system is appropriate basically everywhere that something has to be automatically labelled, identified, registered, stored, monitored or transported. RFID systems are available in a wide scope. Despite the wide range of RFID solutions, each RFID system consists of two components:

a) a transponder and b) a reader [19].

II. MOTIVATION, SCOPE AND LIMITATIONS

RFID systems offer improved efficiency in inventory control, Library Management, Automation systems, logistics and supply chain management. As such, they are of great interest to enterprisers intensively reliant on supply chains, particularly large retailers and consumer manufacturers. We first want to know its various applications in different areas. But without proper protection, wide spread embracing of retail RFID could raise privacy concerns for everyday consumers. The standard of RFID security system is not good enough to protect their system from outside attack. Thus the security issues of RFID are an intriguing research topic. This paper proposes which type of RFID security system is better and highlights the trivial RFID communications.

An organization using the RFID technology has a long term goal to integrate RFID on the retail level for better return on investment (ROI). RFID modernizes the whole software configuration management (SCM) process. And in business the company or organization that has modern and up-to-date supply chain management is expected to be on top of others. On the other hand every educational Institute might build up their library management automation system using RFID Tag. Section III of this paper describe about the basics of RFID system. Here I give a brief idea about RFID components like RFID tags, Readers, Antenna etc., the working technology of RFID system and some possible types of attack in the RFID system. Section IV discusses about the uses and security and privacy issues related to RFID system. In section V, I give some ideas related to using RFID in the coming days and finally conclude the paper in section VI.

III. BACKGROUND STUDY

RFID system is defined by the following three features:

- Electronic identification: The system makes possible an unmistakable labelling of objects by means of electronically stored data.
- Contact less Data transmission: Data identifying the object can be read wirelessly through a radio frequency channel.
- Transmit when requested (on call): A labelled object only transmits data when a matching reader initiates this process. In technical terms, an RFID system consists of two components: a transponder and a reader.

A. Transponders and Readers

The *transponder* also known as a tag acts as the actual data carrier. It is applied to an object (for instance, on a good or package) or integrated into an object (for instance, in a smart card) and can be read without making contact, and rewritten depending on the technology used. Fundamentally the transponder consists of an integrated circuit and a radio-frequency module. An identification number is stored along with other data on the transponder and the object with which it is connected. The reading unit typically called *the reader* as in the following consists of a reading, in some cases a write/read unit and an antenna. The reader reads data from the transponder and in some cases instructs the transponder to store further data. The reader also monitors the quality of data transmission. RFID systems must offer at least the following features:

- Identify the transponder within a specified range.
- Read the data of the transponder.
- Select the transponders relevant for the particular system.
- Guarantee that more than one transponder can be managed within the range of the reader.
- Have some way to recognize errors in order to guarantee operation security.

B. Modes of transmission

Two basically different types of procedure are used to transmit data between the transponder and a reader: duplex procedures including both full duplex (FDX) and half duplex (HDX) and sequential systems (SEQ). The full and half duplex procedures have in common that the energy transmission between reader and transponder is uninterrupted, both in the uplink and in the down link, independently of the data transmission. With sequential systems on the other hand the transponder is supplied with energy only in the pauses in data transmission between the tag and the reader.

RFID systems can be subdivided into three categories by their ranges: close-coupling, remote-coupling and long-range systems. Close-coupling systems have a range up to one centimetre. Close-coupling systems can work with almost any frequencies (from low frequency to 30 MHz), depending on the coupling used. Remote-coupling systems have a range of up to about one meter. They typically work in the frequency range below 135 KHz and at 13.56 MHz. The coupling between the reader and transponder is done inductively. In exceptional cases higher ranges are also possible: 100 meters

or even 1 kilometre, as has been achieved in the frequency spectrum around 5.8 GHz, which is currently in a very early developmental stage.

C. Types of attack

A person who attacks an RFID system may hunt various goals, which can be classified as follows:

- Spying: The attacker gains unauthorized access to information.
- Deception: The attacker deceives the operator or user of an RFID system by feeding in incorrect information.
- Denial of Service (DoS): The availability of functions of the RFID system is compromised.
- Protection of privacy: Because the attacker believes that his privacy is threatened by the RFID system, he protects himself by attacking the system.

And some common security measures are security precautions, authentication, checking the identity of the tag, scrutinizing the identity of the reader, strong mutual authentication, encryption, anti-collision protocols that are safe from eavesdropping, silent tree-walking etc.

IV. RFID USES AND SECURITY ISSUES

I provide some views on security issues concerning RFID systems and highlight some of the areas that have to be considered regarding this topic. To deal with security and RFID means to deal not only with security aspects of RFID systems but also with security aspects of anything or anyone affected by RFID systems. The widespread diffusion of identification technology and storage devices certainly has side effects and can lead to new threats in other areas and applications [17].

A. RFID uses

Access control and personnel tracking and location systems can help to assure the security of restricted areas suppose in airports (such as flight lines, baggage handling areas, customs, employee lounges), passports, for children's security at Schools, Parks, Hospitals and other sensitive areas [9].

1. Passports: RFID tags are used in passports issued by many countries. The first RFID passports (e-passports) were issued by Malaysia in 1998. Malaysian e-passports record the travel history (time, data and Place) of entries and exits from the country.

Now some security concerns on the e-passports. When New Zealand launched e-passport then a source from U.K. mentioned that RFID (radio frequency ID) chips in passports can be cracked in as little as 48 hours [20]. British newspaper The Guardian reports it was able to access the data stored on RFID cards in Britain's newly launched smart passports. However, the New Zealand Department of Internal Affairs (DIA) says there isn't enough information contained within the New Zealand passports' chips to create counterfeit travel documents. DIA passport manager David Philip confirmed that it is possible to access the information stored on the RFID chips and use it to make a clone. However, the RFID chip in

Vol. 8, No. 8, November 2010

the e-passports currently issued in New Zealand is just one security feature out of more than 50 contained in the passport [20]. Having just a cloned chip is not sufficient to create a counterfeit passport, Philip says, and adds that such an end eager is quite involved. While New Zealand passports are "highly desirable," the DIA has seen very few credible counterfeited ones, he says. While the general design goal of the e-passport is to lock the holder's identity to the document in a secure manner, Philip says that there has to be a balance between risk management and customer service [20],[1]. The passport has to be readable around the world in a practical amount of time and preferably in more situations than just immigration. Philip gives airport check-ins as one example of where RFID-equipped passports should be readable. Making the e-passport harder to read is possible, Philp says, but it would make immigration processing take longer and inconvenience people. Researcher Peter Gutmann at the University of Auckland's department of Computer Science is sceptical that the RFID chip provides any real security benefit [16]. In fact, Gutmann goes further and says in his technical background paper, Why biometrics is not a panacea, that RFIDs in passports "are a disaster waiting to happen." German and Dutch passports have already been compromised, according to Gutmann, and this can be done remotely as well [1]. He points to successful attacks by Dutch RFID security specialist Harko Robroch, who intercepted passport and reader device communications from five meters away. Gutmann says eavesdropping on the reader was possible up to 25 meters [20]. In comparison, the Guardian article says U.K. passports are readable 7.5cm away, a far shorter distance than Robroch's interception, but enough in situations such as public transport, where people are close together, to draw off the data stored in the RFID chip.

However, Gutmann's worst-case scenario for RFIDs in passports occurs not when they are being compromised for counterfeiting purposes, but are used to identify the holder. The RFID chip could be used to trigger explosive charges and Gutmann points to a study that shows the current U.S. passport design caused a small, non-lethal explosive charge masked in a rubbish tin to detonate. The New Zealand Department of Internal Affairs (DIA) confirmed reports it is possible to access the information stored on the RFID (radio frequency ID) chips in Britain's newly launched e-passports and use it to make a clone. But said the danger lies not when they are being compromised for counterfeiting purposes, but are used to identify the holder.

Transport payments: Throughout Europe and in Particular in Paris in France (system started in 1995 by the RATP), Lyon and Marseille in France, porto and Lisbon in Portugal, Milan and Torino in Italy, Brussels in Belgium, RFIF passes conforming to the Calypso(RFID) international standard are used for public transport systems. They are also used now in Canada, Mexico, Israel, Bogota and Pereira in Colombia, Scavenger in Norway, etc. Today, the shift to an almost cashless culture has transportation authority's deploying technology that can accelerate the use of wireless

payment and communication systems to support the move towards all electronic open road tolling and emerging traffic management applications, such as high occupancy tolling (HOT) lanes, congestion pricing, dynamic road pricing and express lanes, to mitigate bottleneck congestion or increase infrastructure capacity during peak usage [2].

In this economy, the paper-thin eGo Plus tag, priced under \$10, provides a significant savings for motorists compared to similarly-performing hard case tags that have typically sold for \$25 to \$30 while improving performance capabilities. The sticker tag is comparable in size to a vehicle inspection sticker and mounts easily on a motorist's windshield. The slim form factor also increases point of purchase options making it adaptable to retail outlets and more easily accessible beyond traditional toll customer service centers. Early users of the eGo Plus windshield sticker tag technology experienced two to four times the expected motorist adoption rate, quickly establishing that the paper-thin tag could aid in overcoming deployment barriers that previously hindered widespread motorist use. The paper-thin, battery less ego technology provides environmental value as well.

- Increasing wireless payment of tolls reduces congestion and eliminates idle times at toll plazas, lowering vehicle emissions and improving air quality. By eliminating barriers to adoption, as seen with eGo Plus tags, more motorists will use this form of wireless payment.
- The smaller profile tag consumes less petroleum based raw material to manufacture and reduces transportation and shipping requirements.
- The battery less design of the tag eliminates the additional cost and demand for batteries and subsequent storage and disposal requirements.

The latest in the line of RFID products is the eZGo Anywhere tag. To advance inter operability for electronic toll collection systems nationwide, the eZGo Anywhere standard onboard unit (OBU) is designed to simplify wireless payment of tolls for motorists that travel across states and require different tags for each region's toll system, such as a motorist from the Northeast who travels south to Florida or west to Texas. With these orders, TransCore's eGo family of tags exceeds 11.3 million transponders shipped while globally TransCore's RFID technology deployed in various transportation applications in 41 countries exceeds 35 million RFID tags and 55,000 readers.

- RFID for Library: Among the many uses of RFID technologies is its deployment in libraries. This technology has slowly began to replace the traditional barcodes on library items (books, CDs, DVDs, etc.). Some handy information about RFID for Library is as follows:
 - a. RFID tags replace both the EM security strips and Barcode.
 - b. Simplify patron self check-out / check-in.
 - c. Ability to handle material without exception for video and audio tapes.

- Radio Frequency anti-theft detection is innovative and safe.
- e. High-speed inventory and identify items which are out of proper order.
- f. Long-term development guarantee when using Open Standard.

RFID is the latest technology to be used in library theft detection systems. Unlike EM (Electro-Mechanical) and RF (Radio Frequency) systems, which have been used in libraries for decades, RFID-based systems move beyond security to become tracking systems that combine security with more efficient tracking of materials throughout the library, including easier and faster charge and discharge, inventorying, and materials handling [7].

RFID is a mixture of radio-frequency-based technology and microchip technology. The information contained on microchips in the tags a fixed to library materials is read using radio frequency technology regardless of item orientation or alignment (i.e., the technology does not require line-of-sight or a fixed plane to read tags as do traditional theft detection systems) and distance from the item is not a critical factor except in the case of extra-wide exit gates [14]. The corridors at the building exit(s) can be as wide as four feet because the tags can be read at a distance of up to two feet by each of two parallel exit sensors [14],[8]. The targets used in RFID systems can replace both EM or RF theft detection targets and barcodes.

Advantages of RFID system in library

a. Rapid check-out / check-in- The use of RFID reduces the amount of time required to perform circulation operations. The most significant time savings are attributable to the facts that information can be read from RFID tags much faster than from bar codes and that several items in a stack can be read at the same time [11]. While initially unreliable, the anticollision algorithm that allows an entire stack to be check-out or check-in now appears to be working well. The other time savings realized by circulation staff are modest unless the RFID tags replace both the EM security strips or RF tags of older theft detection systems and the barcodes of the library management system - i.e., the system is a comprehensive RFID system that combines RFID security and the tracking of materials throughout the library; or it is a hybrid system that uses EM for security and RFID for tracking, but handles both simultaneously with a single piece of equipment [14]. There can be as much as a 50 percent increase in throughput. The time savings are less for check-out than for check-in because the time required for check-out usually is extended by social interaction with patrons. For patrons using self check out, there is a marked improvement because they do not have to carefully place materials within a designated template and they can check out several items at the same time. Patron self check-in shifts that work from staff to patrons. Staff is relieved further when readers are installed in book drops [2].

- b. High reliability-
- i. The readers are highly reliable. RFID library systems claim an almost 100 percent detection rate using RFID tags.
- ii. There is no false alarm than with older technologies once an RFID system is properly tuned.
- iii. RFID systems encode the circulation status on the RFID tag. This is done by designating a bit as the "theft" (EAS) bit and turning it off at time of check-out and on at time of check-in. If the material that has not been properly check-out is taken past the exit sensors, an immediate alarm is triggered [3].
 - c. High-speed inventorying -

A unique advantage of RFID systems is their ability to scan books on the shelves without tipping them out or removing them. A hand-held inventory reader can be moved rapidly across a shelf of books to read all of the unique identification information. Using wireless technology, it is possible not only to update the inventory, but also to identify items which are out of proper order [19].

d. Automated materials handling-

Another application of RFID technology is automated materials handling. This includes conveyer and sorting systems that can move library materials and sort them by category into separate bins or onto separate carts. This significantly reduces the amount of staff time required to ready materials for re-shelving.

e. Long tag life-

Finally, RFID tags last longer than barcodes because nothing comes into contact with them. Most RFID vendors claim a minimum of 100,000 transactions before a tag may need to be replaced.

4. RFID for children's security at Schools, Parks, Swimming pools etc.: Traditionally, school facilities have been characterized as easily accessible, open to anyone seeking access. The historical absence of security threats facilitated this culture of openness, which schools have been reluctant to abandon even in the face of changing circumstances like terrorism. Due to the ongoing global terrorism, organized kidnapping of rich people's children and in the face of other social critical issues, the traditional school security systems have proven to be insufficient. The same is applicable in other institution and infrastructure for children's like Park, swimming pools etc. Therefore, RFID being the only technology capable to tracking and identifying any person on the move provides a perfect school and children's security model in the current and future security context. School authorities in the Japanese city of Osaka are now chipping children's clothing, back packs, and student IDs in a primary school [23].

A school in Doncaster, England is piloting a monitoring system designed to keep tabs on pupils by tracking radio chips in their uniforms [21]. St Charles Sixth Form College in West London, England, started September, 2008, is using an RFID card system to check in and out of the main gate, to both track attendance and prevent unauthorized entrance. As is Whitcliffe Mount School in Cleckheaton, England which uses RFID to

payroll system can be managed based on the employee work hour performed as reported by the RFID System.

track pupils and staff in and out of the building via specially designed cards. In the Philippines, some schools already use RFID in IDs for borrowing books and also gates in those particular schools have RFID ID scanners for buying items at a school shop and canteen, library and also to sign in and sign out for student and teacher's attendance. These schools are Claret School of Quezon City, Colegio de San Juan de Letran, San Beda College and other private Schools [21].

Benefits of RFID security at schools

Parent & authority will always be informed about a kid's or student's real time location wherever they are inside the school, park or swimming pool area. School, Park and authority will be notified and or alarmed if any student or kids wants to go away from the authorized area or premise. Students and kid's wristband tag or RFID card will be always visible for visual check by security personnel and can be read by RFID system automatically in the entrance and exit of the school or parks [11]. Parents or authorized guardians will also carry the authorized ID for children's that will authorize them to enter the school or park designated area. School authority will be able to track and monitor attendance of students at School at real time [4]. Parents/guardians can be notified instantly and automatically in case of his/her child goes out of school before scheduled time. It also helps the teachers to record exact time of student's attendants for yearly review of a student's discipline [4].

5. Employee tracking and attendance time: All printed photo IDs are subject to counterfeiting, alteration, duplication, and forgery. Deltech's secured photo ID card gives you highest level of security so you know who's who. Deltech's RFID technology makes photo ID and access control easy while taking security a step further, ensuring that only authorized individuals are able to access your office or secured building and only authorized person/authority can view the details of a person of your organization [4]. Whether you need to identify people or control access, Deltech's ID card gives you the peace of mind that the right people have access to the right places. We can also provide and integrate Biometric finger print or face recognition based access control along with RFID [4].

Benefits of using Deltech RFID Employee tracking Systems:

You can track your employee or staff no matter where they are in the building you can monitor when your staff or employee gets into your office building and when they exit and you can monitor the real time attendance/location of your employee from anywhere in the world [5]. In case of any emergency you will be able to read the employees/staff's urgent information on the move using any mobile reader (optional). I.e. an employee meets an accident during the work or in the area of operation. His/her encoded information in the ID card gives you an easy access to his/her all required information including blood group, address etc without exposing the information to unwanted people [21]. An unauthorized person will not be able to access the restricted areas. Visitors will not be able to leave the premises without returning the visitors card on the security gate/counter. Auto

6. RFID in Hospitals:

Deltech RFID system can be used to track patients, doctors, nurses and expensive equipment in hospitals in real time. RFID tags can be attached to the ID bracelets of all patients, or just patients requiring special attention, so their location can be tracked constantly [4], [5]. Deltech RFID technology can also provide an electronic link for wirelessly communicating patient data. An instant assessment of critical equipment and personnel locations is also possible through RFID technology. Del Technology Limited has implemented RFID solution at Apollo Hospitals Dhaka to be able to identify the location of Doctors, Nurses and other employees and also to be able to register each patients using RFID tags that stores patient's data into the RFID chip. RFID Systems of Deltech implemented at Apollo Hospitals Dhaka, Bangladesh. These applications can be combined with Deltech RFID and or Biometric access control to allow only authorized personnel to access to critical areas of the hospital [6].

Benefits of using Deltech RFID Systems for Hospitals:

- Continuously track each patient's location.
- Track the location of doctors and nurses in the hospital.
- Track the location of expensive and critical instruments and equipment.
- Restrict access to drugs, paediatrics, and other highthreat areas to authorized staff.
- observe and track unofficial persons who are loitering around high-threat areas.
- Facilitate triage processes by restricting access to authorized staff and "approved" patients during medical emergencies, epidemics, terrorist threats, and other times when demands could threaten the hospital's ability to effectively deliver services [6].
- Use the patient's RFID tag to access patient information for review and update through a handheld computer.

7. Animal Identification:

The National Animal Identification System (NAIS) is a government-run system in United States to identify animals and the premises where they have been, in order to provide the potential to identify and isolate threatening diseases. The cattle system is expected to use individual identification with information of the animals' current and previous locations and dates of transfer, sent to a central database. The details of a national plan are still being developed and debated, and changes may occur before finalized. This factsheet is an attempt to help producers understand the NAIS as proposed and interpreted. RFID tags for animals represent one of the oldest uses of RFID technology. Originally meant for large ranches and rough terrain, since the outbreak of Mad Cow Disease, RFID has become crucial in animal identification management. A variety of RFID tags or transponders can also

be used for animal identification. The transponders are more well-known as passive RFID technology, or simply "Chips" on animals [24].

8. Human identification:

The success of various animal identification uses since the early 1990s has spurred RFID research into various human tracking alternatives. Impart-able RFID chips designed for animal tagging are now being used in humans. An early experiment with RFID implants was conducted by British professor of cybernetics Kevin Warwick, who implanted a chip in his arm in 1998. In 2004 Conrod Chase offered implanted chips in his night clubs in Barcelona and Rotterdam to identify their VIP customers, who in turn use it to pay for drinks. In 2004, the Mexican Attorney General's office implanted 18 of its staff members with the Verichip to control access to a secure data room [18].

Security experts have warned against using RFID for authenticating people due to the risk of identity theft. For instance a man-in-the-middle attack would make it possible for an attacker to steal the identity of a person in real-time. Due to the resource constraints of RFIDs it is virtually impossible to protect against such attack models as this would require complex distance-binding protocols [18][15][13][22]. Privacy advocates have protested against Impart-able RFID chips, warning of potential abuse and denouncing these types of RFID devices as "spychips," and that use by governments could lead to an increased loss of civil liberties and would lend itself too easily to abuse. One such case of this abuse would be in the microchip's dual use as a tracking device. Such concerns were justified in the United States, when the FBI program COINTELPRO was revealed to have tracked the activities of high profile political activist and dissident figures.

There is also the possibility that the chip's information will be available to those other than governments, such as private business, thus giving employers highly delicate information about employees. In addition, privacy advocates state that the information contained in this chip could easily be stolen, so that storing anything private in it would be to risk identity theft. According to the FDA, implantation of an RFID chip poses potential medical downsides. Electrical hazards, MRI incompatibility, adverse tissue reaction, and migration of the implanted transponder are just a few of the potential risks associated with the Verichip ID implant device, according to an October 12, 2004 letter issued by the Food and Drug Administration (FDA).

9. RFID in inventory system:

An advanced automatic identification technology such as the Auto-ID based on the RFID technology as two values for inventory systems is already in use. First, the visibility provided by this technology allows an accurate knowledge on the inventory level by eliminating the discrepancy between inventory record and physical inventory. In an academic study performed at Wal-Mart, RFID reduced out of Stocks by 30 percent for products selling between 0.1 and 15 units a day. Second, the RFID technology can prevent or reduce the

sources of errors. Benefits of using RFID include the reduction of labour costs, the simplification of business processes and the reduction of inventory inaccuracies. Wal-Mart and the United States Department of Defence have published requirements that their vendors placed RFID tags on all shipments to improve supply chain management.

10. Other RFID Uses

NADRA (National Database and Registration Authority) in Pakistan has developed an RFID-based driving license that has bears the license holders personal information and stores data regarding traffic violations, tickets issued and outstanding penalties. The license cards are designed so that driving rights can be evoked electronically in case of serious violations. Sensors such as seismic sensors may be read using RFID transceivers, greatly simplifying remote data collection. In august 2004, the Ohio Department of Rehabilitation and Correction (ODRH) approved a \$415000 contract to evaluate the personal tracking technology of Alanco Technologies. Inmates will wear wristwatch-sized transmitters that can detect attempted removal and alert prison computers. Facilities in Michigan, California and Illinois already employ the technology. RFID in designed by Vita Craft, is an automatic cooking device that has three different sized pans, a portable induction heater and recipe cards. Each pan is embedded with a RFID tag that monitors the food 16 times per second while a MI tag in the handle of the pans transmits signals to the induction heater to adjust the temperature.

B. Security issues

Some problems with RFID are reported during the use. RFID problems can be divided into several categories:

- Technical problems with RFID
- Privacy and ethics problems with RFID.

Technical problems with RFID

RFID has been implemented in different ways by different manufacturers; global standards are still being worked on. It should be noted that some RFID devices are never meant to leave their network (as in the case of RFID tags used for inventory control within a company). This can cause problems for companies. Consumers may also have problems with RFID standards. For example, ExxonMobil's Speed Pass system is a proprietary RFID system; if another company wanted to use the convenient Speed Pass (say, at the drive-in window of your favourite fast food restaurant) they would have to pay to access it - an unlikely scenario. On the other hand, if every company had their own "Speed Pass" system, a consumer would need to carry many different devices with them.

An RFID system can utilize a few standards. The problem has been that there is no one universally accepted standard. Competing standards have been one of the more difficult issues for RFID, and as a result, most RFID applications have been closed systems. Standards and specifications may be set at the international, national, industry or trade association level, and individual organizations may term their own

specifications as standard. Many industry standards and specifications set by individual organizations are based on international standards to make implementation and support easier and to provide a wider choice of available products. Standards can be applied to include the format and content of the codes placed on the tags, the protocols and frequencies that will be used by the tags and readers to transmit the data, the security and tamper-resistance 52 of tags on packaging and freight containers, and applications use. The two largest drivers for RFID today are Wal-Mart and the Department of Defense (DOD). Both have issued mandates for their top suppliers to use RFID technology when shipping products to their distribution centers. They are both looking to accomplish the same thing, but have a slightly different long-term outlook [10].

The ISO (International Standards Organization) and the EPC (Electronic Product Code) Global have both been leading figures in this debate. The ISO has their 18000 standard and the EPC Global Center has introduced the EPC standard. Wal-Mart has decided to use the EPC standard, where the DOD wants to use the EPC for general purposes, but use the ISO standard for air interface. This is putting a lot of pressure on the ISO and EPC to come to some kind of an agreement. EPC standard for air interface is not compatible with the ISO 18000 UHF (Part 6) standard. Both, the EPC and ISO 18000 (Part6) standards, deal with the tracking of merchandiser through the supply chain. This is WalMarts and the Department of Defenses primary focus at this time. The ISO 18000 (Part 6) standard only deals with air interface protocols, whereas the EPC standard also includes data structure. The desire is for these two protocols not to be mutually exclusive. There are several evolutions to the EPC standard. Class 1-Generation 1 is the current version of EPC. It is not backward compatible with Class 0. Generation 2 was hoped to be backward compatible with Class 0 but merging with the ISO 18000 standard will be difficult, if not impossible. Wal-Mart has said it will support both Class 0 and 1 but wants to settle on Class 1 Generation 2 when it is finalized. The EPC standard was originally developed for carton and pallet tracking within the supply chain.

Privacy and ethics problems with RFID

The following problems with RFID tags and readers have been reported.

- The contents of an RFID tag can be read after the item leaves the supply chain.
- An RFID tag cannot tell the difference between one reader and another. RFID scanners are very portable; RFID tags can be read from a distance, from a few inches to a few yards. This allows anyone to see the contents of your purse or pocket as you walk down the street. Some tags can be turned off when the item has left the supply chain.
- RFID tags are difficult to remove RFID tags are difficult to for consumers to remove; some are very small (less than a half-millimetre square, and as thin as a sheet of paper) others may be hidden or embedded inside a product where consumers cannot see them. New technologies allow RFID

tags to be "printed" right on a product and may not be removable at all.

- RFID tags can be read without your knowledge Since the tags can be read without being swiped or obviously scanned (as is the case with magnetic strips or barcodes), anyone with an RFID tag reader can read the tags embedded in your clothes and other consumer products without your knowledge. For example, you could be scanned before you enter the store, just to see what you are carrying. You might then be approached by a clerk who knows what you have in your backpack or purse, and can suggest accessories or other items.
- RFID tags can be read a greater distances with a high-gain antenna For various reasons, RFID reader/tag systems are designed so that distance between the tag and the reader is kept to a minimum (see the material on tag collision above). However, a high-gain antenna can be used to read the tags from much further away, leading to privacy problems.
- RFID tags with unique serial numbers could be linked to an individual credit card number. At present, the Universal Product Code (UPC) implemented with barcodes allows each product sold in a store to have a unique number that identifies that product. Work is proceeding on a global system of product identification that would allow each individual item to have its own number. When the item is scanned for purchase and is paid for, the RFID tag number for a particular item can be associated with a credit card number.

Authentication

When authentication is carried out, the identity of a person or a program is checked. Then, on that basis, authorization takes place, i.e. rights, such as the right of access to data, are granted. In the case of RFID systems, it is particularly important for tags to be authenticated by the reader and vice-versa. In addition, readers must also authenticate themselves to the backend, but in this case there are no RFID-specific security problems. Checking the Identity of the tag When the RFID system detects a tag, it must check its identity in order to ascertain if the tag has the right to be part of the system at all.

A worldwide and unambiguous regulation for issuing ID numbers, as proposed, for example, in the form of the Electronic Product Code (EPC), offers a certain amount of protection from falsified tags. At the very least, the appearance of numbers that were never issued or of duplicates (cloning) can be recognized in certain applications. In addition, authentication may take place via the challenge-response system, in which the reader sends a random number or a time stamp to the tag (challenge) which the tag returns in encrypted form to the reader (response). The key used in this case is a jointly known secret by means of which the tag proves its identity. The decisive element in this procedure is the fact that the key itself is never transmitted and that a different random number is used for every challenge. As a result, the reader cannot be deceived by the communication being recorded and replayed (replay attack). This unilateral authentication procedure is defined as a "symmetric-key two-pass unilateral

Vol. 8, No. 8, November 2010

authentication protocol" in ISO Standard 9798. An attacker would have to get hold of the key which is stored both on the tag and in the back end of the RFID system. In order to do so, it would be necessary to decode the response data that were transmitted in encrypted form, which is a very complex if not almost impossible task, depending on the length of the key. In principle, the key could also be read by physical means from the storage cells of the chip, but this would require very complicated laboratory methods, such as the "Focused Ion Beam" (FIB) technique. In this procedure, an ion beam removes very thin layers (a few layers of atoms) in separate steps so that the contents can be analyzed microscopically.

C. Data protection and privacy

The progressive implementation of RFID systems is being keenly followed by the public and mass media and is a topic of controversial discussion. From a social point of view, guarantees of privacy and various aspects of data protection play an ever-increasing role in this controversy (catch-words: naked customer or naked citizen). Civil rights organizations have published a common position paper on the use of RFID and the associated risks posed to data privacy. The signatory organizations acknowledged that there can be justified interests in the use of RFID on the part of business but, in light of the considerable risks involved, they called for dealers and manufacturers to observe a voluntary moratorium on the use of RFID for consumer goods until all risks were reviewed in a comprehensive technology assessment that would propose possible counter strategies.

Anyone with an appropriately equipped scanner and close access to the RFID device can activate it and read its contents. Obviously, some concerns are greater than others. If someone walks by your bag of books from the bookstore with a 13.56 MHz "sniffer" with an RF field that will activate the RFID devices in the books you bought, that person can get a complete list of what you just bought. That is certainly an invasion of your privacy, but it could be worse. Another scenario involves a military situation in which the other side scans vehicles going by, looking for tags that are associated with items that only high-ranking officers can have, and targeting accordingly. Companies are more concerned with the increasing use of RFID devices in company badges.

Along with privacy, consumers want a complex and constantly shifting mix of low prices, convenience, customization, quality, customer service, and other characteristics in their goods and services. Radio frequency identification technology will help producers, marketers, and retailers better understand and serve the mix of interests consumers have. The components that go into RFID readers and tags are simple radio communications, but their smaller size and broad deployment enhance the power of the technology and raise concerns about the privacy effects of RFID deployment. These concerns are often premised on unlikely assumptions about where the technology will go and how it will be used. Any inclination to abuse RFID technology will be hemmed in by a variety of social forces, economic forces being one of the most significant. The typical RFID tag

in the consumer goods environment will be cheap, dumb, and not good for much more than tracking inventory. Consumers, as economic actors, have substantial power to dictate in the give and take of the market how RFID will be used. They will likely demand tags linking to their identities in certain applications such as consumer electronics but may object to the presence of RFID tags in other situations. They may demand peel-off tags, or assurances about what a particular tag is doing. In many instances, they will be indifferent, and rationally so. Regulators, think-tank analysts, and activists should not attempt to dictate RFID policy before real experience has been gained and must not set up moratorium on RFID deployment.

D. Comparison with Barcode

Advantages of RFID Versus Barcodes RFID tags and barcodes both carry information about products. However, there are important differences between these two technologies. Barcode readers require a direct line of sight to the printed barcode; RFID readers do not require a direct line of sight to either active RFID tags or passive RFID tags. RFID tags can be read at much greater distances; an RFID reader can pull information from a tag at distances up to 300 feet. The range to read a barcode is much less, typically no more than fifteen feet. RFID readers can interrogate, or read, RFID tags much faster; read rates of forty or more tags per second are possible. Reading barcodes is much more time-consuming; due to the fact that a direct line of sight is required, if the items are not properly oriented to the reader it may take seconds to read an individual tag. Barcode readers usually take a halfsecond or more to successfully complete a read.

Line of sight requirements also limit the ruggedness of barcodes as well as the reusability of barcodes. Since line of sight is required for barcodes, the printed barcode must be exposed on the outside of the product, where it is subject to greater wear and tear. RFID tags are typically more rugged, since the electronic components are better protected in a plastic cover. RFID tags can also be implanted within the product itself, guaranteeing greater ruggedness and reusability. Barcodes have no read/write capability; that is, you cannot add to the information written on a printed barcode. RFID tags, however, can be read/write devices; the RFID reader can communicate with the tag, and alter as much of the information as the tag design will allow. RFID tags are typically more expensive than barcodes, in some cases, much more so [12]. RFID and barcodes are similar in that they are both data collection technologies. However, they differ significantly in many areas. Although this comparison primarily focuses on the advantages of RFID over barcodes, RFID will probably not completely replace barcode technology. Barcodes offer some advantages over RFID, most notably the low cost. A tabular comparison for RFID and Barcode is given in Table I-

TABLE I. COMPARISON OF RFID AND BARCODE

	RFID	Barcode
Read range	Passive RFID:	Several inches up

	Up to 40 feet (fixed	to
	readers).	several feet.
	Up to 20 feet (handheld	
	readers)	
	Active RFID:	
	Up to 100's	
	of feet or more.	
Line of sight	Not required	Required.
	(in most cases).	
Type of	Can uniquely identify	Can typically only
identification	each item/asset tagged.	identify the type of
		item (UPC Code)
		but not uniquely.
Read/Write	Many RFID tags are	Read only.
	Read/Write.	
Read rate	10's, 100's or 1000's	Only one at a time.
	simultaneously.	
Technology	RF (Radio Frequency).	Optical (Laser).
Interference	Like the TSA	Obstructed
	(Transportation	barcodes cannot be
	Security	read (dirt covering
	Administration), some	barcode, torn
	RFID frequencies do	Barcode etc.).
	not like Metal and	
	Liquids. They can	
	cause interference with	
	certain RF Frequencies.	
Automation	Most "fixed" readers do	Most barcode
	not require human	scanners require a
	involvement to collect	human to operate
	data (automated).	(labour intensive).

V. THOUGHT OF NEXT GENERATION

Some vendors have been combining RFID tags with sensors of different kinds. This would allow the tag to report not simply the same information over and over, but identifying information along with current data picked up by the sensor. For example, an RFID tag attached to a leg of lamb could report on the temperature readings of the past 24 hours, to ensure that the meat was properly kept cool. Over time, the proportion of "scan-it-yourself" aisles in retail stores will increase. Eventually, we may wind up with stores that have mostly "scan-it-yourself" aisles and only a few checkout stations for people who are disabled or unwilling [12].

RFID tags come in a wide variety of shapes and sizes; they may be encased in a variety of materials: Animal tracking tags, inserted beneath the skin, can be rice-sized; Tags can be screw-shaped to identify trees or wooden items; Credit-card shaped for use in access applications. The antitheft hard plastic tags attached to merchandiser in stores are also RFID tags. Heavy-duty 120 by 100 by 50 millimetre rectangular transponders are used to track shipping containers, or heavy machinery, trucks, and railroad cars. Many musical instruments are stolen every year. For example, custom-built or vintage guitars are worth as much as \$50,000 each. Snagg, a California company specializing in RFID microchips for

instruments, has embedded tiny chips in 30,000 Fender guitars already [23].

The smallest tags that will likely be used for consumer items do not have enough computing power to do data encryption to protect your privacy. The most they can do is PIN-style or password-based protection [20]. Civil liberties groups (among others) have become increasingly concerned about the use of RFIDs to track the movements of individuals. For example, passports will soon be required to contain some sort of RFID device to speed border crossings. Scanners placed throughout an airport, for example, could track the location of every passport over time, from the moment you left the parking lot to the moment you got on your plane. In June, the Japanese government passed a draft RFID Privacy Guideline that stated the following: " Indication that RFID tags exist "Consumers right of choice regarding reading tags "Sharing information about social benefits of RFID, etc. "Issues on linking information on tags and databases that store privacy information." Restrictions of information gathering and uses when private information is stored on tags " Assuring accuracy of information when private information is stored on tags "Information administrators should be encouraged" Information sharing and explanation for consumers.

There was a recent report revealing clandestine tests at a Wal-Mart store where RFID tags were inserted in packages of lipstick, with scanners hidden on nearby shelves. When a customer picked up a lipstick and put it in her cart, the movement of the tag was registered by the scanners, which triggered surveillance cameras. This allowed researchers 750 miles away to watch those consumers as they walked through the store, looking for related items. Contact less Credit Card Advantages Credit card companies are claiming the following advantages for contact less credit cards: The card is faster to use. To make a purchase, the card owner just waves his card over the RFID reader, waits for the acceptance indicator - and goes on his way. American Express, Visa and Master card have all agreed to waive the signature requirement for contact less credit card transactions under \$25. If we want to look at the numbers, here is where this technology is taking us in our need for speed (average transaction speeds):

- 1. Contact less credit card transaction: 15 seconds
- 2. Magnetic strip card transaction: 25 seconds
- 3. Cash transaction: 34 seconds

The contact less cards use highly secure data transmission standards. Contact less cards make use of the most secure encryption standards practical with current technology. 128-bit and triple DES encryption make it nearly impossible for thieves to steal your data.

Contact less Credit Card Disadvantages

Contact less cards are more exposed than regular credit cards. If you want to keep your credit card secure, you could keep it safely in an enclosed wallet or purse; thieves would have absolutely no way to even know if you have a credit card. However, a thief armed with a suitable reader, within a few feet of you, would be able to interrogate all of the cards in

your wallet or purse without your knowledge. Also, a regular credit card transaction is fairly secure; the magnetic strip is swiped at very close range (less than a millimetre). However, a thief with a suitable reader could monitor your contactless card transaction while standing at the counter with you, or just behind you. These concerns have, of course, been carefully noted by credit card companies. The RFID chip in the contact less credit card responds to the merchant reader with a unique number used for that transaction only; it does not simply transmit the consumer's account number. This number is also encrypted. It is easier to spend. Studies have demonstrated that consumers will be more likely to spend, and will spend more frequently, with contact less credit cards [12].

VI. CONCLUSION

Finally I would like to conclude by mentioning some future challenges regarding RFID technology. Challenges will arise from the flexibility of changes in tag ownership. Today, domain names, for example, do not change hands very often; the DNS can involve human intermediated access-control. Another important aspect of RFID security is that of user perception of security and privacy in RFID systems. As users cannot see RF emissions, they form their impressions based on physical cues and industry explanations. RFID will come to secure ever more varied forms of physical access and logical access. Every technology has some advantages and disadvantages; but RFID technology has so far showed a lot of potential to be a topic on which intense research can be carried upon.

REFERENCES

- [1] http://rfid.weblogsinc.com, accessed on July 2009.
- [2] http://rfid.weblogsinc.com,accessed on January 2010.
- [3] http://www.deltechbd.com/advantage_rfid.php, accessed on April 2010.
- [4] http://www.deltechbd.com/advantage_rfid.php, accessed on January 2010.
- [5] http://www.deltechbd.com/imanage accesscontrol.php, accessed on February 2010.
- [6] http://www.rfid-industry.com/ar/9b.htm, accessed on June 2010.
- [7] http://www.rfid-library.com/en/rfid-transponder.html, accessed on December 2009.
- [8] http://www.rfid-library.com/en/system-flash-demo.html, accessed on April 2010.
- http://www.rfidjournal.com/, accessed on December 2009.
- [10] www.scansource.eu/en/education.htm?eid=12&elang=en, accessed on March 2010.
- [11] http://www.technovelgy.com/ct/Technology-Article.asp?ArtNum=20,accessed on April 2010.
- [12]http://www.technovelgy.com/ct/Technology-Article.asp? accessed on April 2010.
- [13] EPCglobal Inc. EPCTM"Generation 1 tag data standards version 1.1 rev. 1.27". http://www.verisign.com/static/015884.pdf, May 2005.
- [14]http://www.softlinkasia.com/RFID.htm, accessed on April

- [15] Epcglobal Inc. EPCTM generation 1 tag data standards version 1.1 rev.http://www.epcglobalinc.org. 2005.
- [16] Machine readable travel documents, Part 1: Machine Readable Passports, Volume 1: Passports with machine readable data stored in optical character recognition format, Sixth edition 2006, International Civil Aviation Organization.
- [17] Oertel, B., Wolk, M., Hilty, L., Kohler, A., Kelter, H., Ullmann, M., et al. (2004). Security Aspects and Prospective Applications of RFID Systems. Retrieved On 08/01/2006 from www.bsi.de/fachthem/rfid/RIKCHA englisch.pdf.
- [18] Jonathan Collins, Tag Encryption for Libraries To protect patrons' privacy, a new system encrypts data stored on a book's RFID tag, retrieved on June 2010 from http://www.rfidjournal.com/article/pdf/1027 /1/1/rfidjournal-article1027.PDF.
- [19] K. Fishkin and J. Lundell. RFID in healthcare. In S. Garfinkel and B. Rosenberg, editors, RFID: Applications, Security, and Privacy, pages 211-228, Addison-Wesley, 2005.
- [20] Juha Saarinen, Computerworld-New Zealand, 2006 accessed on June, 2010 from http://computerworld.co.nz/news.nsf/NL/61DD9AC9B0 1A7D68CC257230000696BD.
- [21] G.P Hancke and M.G. Kuhn, An RFID distance bounding protocol, Conference on Security and Privacy in Communication Networks (SECURECOMM 2005), pp 67-73, September 2005.
- [22] S. Inoue and H. Yasuura, RFID privacy using usercontrollable uniqueness. In RFID Privacy Workshop. MIT, November 2003.
- [23] Arie Jules, RFID Security and Privacy: A Research Survey, IEEE Journal on Selected Areas in Communication, Vol. 24, No. 2, February 2006.
- [24] Claire McEntee, 'Old technology' for \$23m cattle tracing scheme - Low frequency RFID plan challenged, Computerworld-Newzealand, 2009, accessed from http://computerworld.co.nz/news.nsf/tech/2E80078E34 C21EC5CC25768B006CBA36.

AUTHORS PROFILE

Mohammad Tauhidul Islam received his bachelor's degree from Islamic University of Technology, Gazipur, Bangladesh in 2005 and M.Sc. from University of Lethbridge, Alberta, Canada in 2009. His research interest includes wireless sensor networks, security issues related to RFID and study and analysis of hard problems and possible approximation algorithms for those.

An Improved Fuzzy Time Series Model For Forecasting

Ashraf K. Abd-Elaal¹

Department of Computer and Information Sciences The High Institute of Computer Science Sohag, Egypt

Hesham A. Hefny

Department of Computer and Information
Sciences,
Institute of Statistical Studies and Research,
Cairo University, Egypt

Ashraf H. Abd-Elwahab

Department of Computer Sciences
Electronics Research Institute
National Center for Research
Cairo, Egypt

Abstract— Researchers introduce in this paper, an efficient fuzzy time series forecasting model based on fuzzy clustering to handle forecasting problems and improving forecasting accuracy. Each value (observation) is represented by a fuzzy set. The transition between consecutive values is taken into account in order to model the time series data. Proposed model employed eight main steps in time-invariant fuzzy time-series and time-variant fuzzy time series models to increase the performance of the proposed fuzzy time series model. The method of FCMI is integrated in the processes of fuzzy time series to partition datasets. The proposed model has been implemented to forecast the world production of iron and steel and the enrollments of the University of Alabama. The proposed model provide higher accuracy in forecasting. Our results show that this approach can lead to satisfactory performance for fuzzy time series

Keywords-forecasting; fuzzy Clustering; fuzzy time series; iron.

I. INTRODUCTION

Traditional forecasting methods can deal with many forecasting cases, but they cannot solve forecasting problems in which the historical data are linguistic values. Song and Chissom [12] presented the concept of fuzzy time series based on the historical enrollments of the University of Alabama. They presented the time-invariant fuzzy time series model and the time-variant fuzzy time series model based on the fuzzy set theory for forecasting the enrollments of the University of Alabama.

The fuzzy forecasting methods can forecast the data with linguistic values. Fuzzy time series do not need to turn a non-stationary series into a stationary series and do not require more historical data along with some assumptions like normality postulates. Although fuzzy forecasting methods are suitable for incomplete data situations, their performance is not always satisfactory [9,11].

Huarng [6] proposed heuristic models; by integrating problem-specific heuristic knowledge to improve forecasting.

¹ Corresponding Author: Ashraf K. Abd-Elaal

Tsaur, et al [14] proposed an analytical approach to find the steady state of fuzzy relation matrix to revise the logic forecasting process. Based on the concept of fuzziness in Information Theory, the concept of entropy is applied to measure the degrees of fuzziness when a time-invariant relation matrix is derived. In order to show the forecasting performance, the best fitted regression equations are applied to compare with the proposed method.

Yu [15] proposed weighted models to tackle two issues in fuzzy time series forecasting; namely, recurrence and weighting. Weighted fuzzy time series models appear quite similar to the weight functions in local regression models; however, both are different. The local regression models focus on fitting using a small portion of the data, while the fuzzy relationships in weighted fuzzy time series models are established using the possible data from the whole of the database.

Jilani and Burney [7] presented two new multivariate fuzzy time series forecasting methods. These methods assume m-factors with one main factor of interest. Stochastic fuzzy dependence of order k is assumed to define general methods of multivariate fuzzy time series forecasting and control.

Cheng et al [4] proposed a novel multiple-attribute fuzzy time series method based on fuzzy clustering. The methods of fuzzy clustering were integrated in the processes of fuzzy time series to partition datasets objectively and enable processing of multiple attributes.

Abd Elaal et al [1-2] proposed a novel forecasting fuzzy time series model depend on fuzzy clustering for improving forecasting accuracy. Kai et al [8] proposed a novel forecasting model for fuzzy time series using K-means clustering algorithm for forecasting.

In this paper, researchers propose an efficient fuzzy time series forecasting model based on fuzzy clustering to handle forecasting problems and improving forecasting accuracy. Each value (observation) is represented by a fuzzy set. The transition between consecutive values is taken into account in order to model the time series data.

II. RELATED WORKS

In this section, two related works including: fuzzy clustering and fuzzy time series.

A. Fuzzy clustering (FCMI)

Fuzzy c-means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. Fuzzy C-Mean Iterative assume that: the existence of pattern space $X=\{x_1, x_2,..., x_m\}$ and c fuzzy clusters, whose centers have initial values $y_{10}, y_{20},...,y_{c0}$. Every iteration the membership function values updated and the cluster centers also. The process terminates when the difference between two consecutive clusters centers do not exceed a given tolerance [5].

$$d_{ij}^{(k)} = \left\| x_j - y_i^{(k)} \right\| \tag{1}$$

Fuzzy clustering is carried out through an iterative optimization of the objective function d_{ij} , with the update of membership u_{ij} and the cluster centers y_i by:

$$u_{ij}^{(k)} = \left[\sum_{l=1}^{c} \left(\frac{d_{ij}^{(k)}}{d_{ij}^{(k)}} \right)^{2/(\beta-1)} \right]^{-1}$$
 (2)

$$y_{i}^{(k+1)} = \frac{\sum_{j=1}^{m} u_{ij}^{(k)} x_{j}}{\sum_{j=1}^{m} u_{ij}^{(k)}}$$
(3)

This iteration will stop when

$$\left[\sum_{i=1}^{c} \left\| y_{i}^{(k+1)} - y_{i}^{(k)} \right\|^{2} \right]^{1/2} < \varepsilon \tag{4}$$

B. Fuzzy time series

Song and Chissom [13] presented the concept of fuzzy time series based on the historical enrollments of the University of Alabama. Fuzzy time series used to handle forecasting problems. They presented the time-invariant fuzzy time series model and the time-variant fuzzy time series model based on the fuzzy set theory for forecasting the enrollments of the University of Alabama. The definitions and processes of the fuzzy time-series presented by Song and Chissom are described as follows [6,12].

Definition 1. (FTS) Assume Y (t) (t = ...0, 1, 2, ...) is a subset of a real numbers. Let Y (t) be the universe of discourse

defined by the fuzzy set $f_i(t)$. If F(t) is a collection of $f_1(t)$, $f_2(t)$. . . then F(t) is defined as a fuzzy time-series on Y(t) ($t = \ldots, 0, 1, 2, \ldots$).

Definition 2. (FTSRs) If there exists a fuzzy logical relationship R(t-1, t), such that $F(t) = F(t-1) \times R(t-1, t)$, where "×" represents an operation, then F(t) is said to be induced by F(t-1). The logical relationship between F(t) and F(t-1) is $F(t-1) \rightarrow F(t)$.

Definition 3. (FLR) suppose $F(t-1) = A_i$ and $F(t) = A_j$. The relationship between two consecutive observations, F(t) and F(t-1), referred to as a fuzzy logical relationship, can be denoted by $A_i \rightarrow A_j$, where Ai is called the Left-Hand Side (LHS) and A_i the Right-Hand Side (RHS) of the FLR.

Definition 4. (FLRG) All fuzzy logical relationships in the training dataset can be grouped together into different fuzzy logical relationship groups according to the same Left-Hand Sides of the fuzzy logical relationship. For example, there are two fuzzy logical relationships with the same Left-Hand Side $(A_i): A_i \rightarrow A_{j1}$ and $A_i \rightarrow A_{j2}$. These two fuzzy logical relationships can be grouped into a fuzzy logical relationship group $A_i \rightarrow A_{j1}$ A_{j2} .

Definition 5. (IFTS & VFTS) Assume that F(t) is a fuzzy time-series and F(t) is caused by F(t-1) only, and $F(t) = F(t-1) \times R(t-1,t)$. For any t, if R(t-1,t) is independent of t, then F(t) is named a time-invariant fuzzy time-series, otherwise a time-variant fuzzy time-series.

a) Song and Chissom model

Song and Chissom employed five main steps in time-invariant fuzzy time-series and time-variant fuzzy time series models as follows:

Step 1: Define the universe of discourse U. Define the universe of discourse for the observations. According to the issue domain, the universe of discourse for observations is defined as,

$$U = [D_{min} - D_1, D_{max} + D_2]$$
 (5)

where, D_{min} is the minimum value, D_{max} is the maximum value, D_1 , D_2 is the positive real numbers.

Step 2: Partition universal of discourse U into equal intervals.

Step 3: Define the linguistic terms. Each linguistic observation, A_k can be defined by the intervals $u_1, u_2, ..., u_n$, as follows:

$$\mathbf{A}_{k} = \begin{cases} \frac{\frac{1}{u_{1}} + \frac{0.5}{u_{2}}}{\frac{0.5}{u_{k-1}} + \frac{1}{u_{k}} + \frac{0.5}{u_{k+1}}} & 2 \le k \le n-1\\ \frac{\frac{0.5}{u_{n-1}} + \frac{1}{u_{n}}}{\frac{0.5}{u_{n-1}} + \frac{1}{u_{n}}} & k=n \end{cases}$$
(6)

Step 4: Fuzzify the historical data. Each historical data can be fuzzified into a fuzzy set.

Step 5: Build fuzzy logic relationships. Build fuzzy logic relationships. Two consecutive fuzzy sets $A_i(t-1)$ and $A_j(t)$ can be established into a single FLR as $A_i \rightarrow A_j$.

III. PROPOSED MODEL

In this section we proposed an efficient fuzzy time series forecasting model based on fuzzy clustering to handle forecasting problems and improving forecasting accuracy. Most researchers have been taken the same way according to processes of the fuzzy time-series, which are presented by Song and Chissom, but we introduce a novel model based on fuzzy clustering to determine the membership values not as Song and Chissom model, and to increase the performance. Proposed model employed eight main steps in time-invariant fuzzy time-series and time-variant fuzzy time series models as follows:

Step 1: Cluster data into c clusters: Apply fuzzy clustering on a time series Y(t) with n observation to cluster this time series into c ($2 \le c \le n$) clusters. FCMI is used because it is the most popular one and well known in fuzzy clustering field.

Step 2: Determine membership values for each cluster: In this step, membership values is determining after doing fuzzy cluster. The proposed model selected the maximum membership grade of each value for each cluster which it belong to.

Step 3: Rank each cluster: Proposed model ranking clusters by the center of each cluster, where first cluster has the minimum center, and last cluster has the maximum center.

Step 4: Define the universe of discourse U: In this step, the proposed model defines the universe of discourse as Song and Chissom were defined it as in (5).

Step 5: Partition universal of discourse U into equal intervals: According to this step, the proposed model, partition the universe of discourse into c intervals.

Step 6: Fuzzify the historical data: In this step, proposed model fuzzufy historical data, where the proposed model determine the best fuzzy cluster to each actual data

Step 7: Build fuzzy logic relationships: Proposed model in this step build fuzzy logic relationship as definition 3. if F(t-1) = Ai and F(t) = Aj then the relationship between two consecutive observations: Ai \rightarrow Aj

Step 8: Calculate forecasting outputs: The forecasting value for each cluster is calculated by proposed model as:

$$forecaste\left(A_{i}\right) = \frac{df_{1} \times X_{1} + df_{1} \times X_{1} + \dots + df_{m} \times X_{m}}{\sum\limits_{i=1}^{m} df_{j}} \tag{7}$$

Where dfj is the membership grade,

Xi is the actual value.

A. Evaluating of the proposed model

To evaluating the performance of the proposed model, the researchers compare the forecasting values of enrollments of the University of Alabama with some famous models such as Jilani and Burney [7], Tsaur and Yang [14], Yu [15], Kai et al [8], and Cheng, et al [4].

The forecasting accuracy is compared by using (NRMSE) Normalized Root Mean Square Error. NRMSE, in statistic is the square root of the sum of the squared deviations between actual and predicted values divided by the sum of the square of actual values.

$$NRMSE = \sqrt{\frac{\sum_{i=1}^{N} (actual_i - predict_i)^2}{\sum_{i=1}^{N} (actual_i)^2}}$$
(8)

In this study, to evaluate the forecasting accuracy of the proposed model, the researchers use the enrollments of the University of Alabama as the forecasting target in the existing forecasting models.

Based on the enrollments of the University of Alabama from 1971 to 1992, we can get the universe of discourse U=[13055,19337], partition U into 7 equal intervals, D_1 =13, and D_2 =55. Hence, the intervals are u_1 ; u_2 ; u_3 ; u_4 ; u_5 ; u_6 ; u_7 ; where :-

$$u_1$$
=[13024.00, 13933.71]
 u_2 =[13933.71, 14843.43],
 u_3 =[14843.43, 15753.14],
 u_4 =[15753.14, 16662.86],
 u_5 =[16662.86, 17572.57],
 u_6 =[17572.57, 18482.29],
 u_7 =[18482.29, 19392.00],

Table I lists the enrollment of the University of Alabama from 1971 to 1992, and membership grades of enrollments for each linguistic. Define the fuzzy set A_i using the linguistic variable "Enrollments of the University of Alabama", let A_1 = (very very few), A_2 = (very few), A_3 = (few), A_4 = (moderate), A_5 = (many), A_6 = (many many), A_7 = (too many). The proposed model selected the maximum membership grade for each cluster, the forecasting value for each cluster calculating as in (7):

forecaste
$$(A_1) = \frac{1 \times (1972)}{1} = 13563$$

forecaste $(A_2) = \frac{0.8 \times (1984)}{0.8} = 15145$

forecaste
$$(A_3) = \frac{1 \times (1975) + 1 \times (1982)}{2} = 15446$$

forecaste $(A_4) = \frac{1 \times (1978)}{1} = 15861$
forecaste $(A_5) = \frac{1 \times (1979)}{1} = 16807$

forecaste
$$(A_6) = \frac{1 \text{ x (1988)}}{1} = 18150$$

forecaste
$$(A_7) = \frac{1 \times (1989)}{1} = 18970$$

TABLE I. DATA OF ENROLLMENTS OF THE UNIVERSITY OF ALABAMA AND MEMBERSHIP GRADES.

Year	Actual	$\mathbf{A_1}$	A ₂	A_3	A_4	A ₅	A ₆	A ₇
1971	enrollm 13055	0.8	0.1	0	0	0	0	0
1972	13563	1	0	0	0	0	0	0
1973	13867	0.9	0.1	0	0	0	0	0
1974	14696	0.1	0.7	0.2	0.1	0	0	0
1975	15460	0	0	1	0	0	0	0
1976	15311	0	0.1	0.9	0	0	0	0
1977	15603	0	0.1	0.6	0.3	0	0	0
1978	15861	0	0	0	1	0	0	0
1979	16807	0	0	0	0	1	0	0
1980	16919	0	0	0	0	0.9	0	0
1981	16388	0	0	0.1	0.3	0.6	0	0
1982	15433	0	0	1	0	0	0	0
1983	15497	0	0	0.9	0.1	0	0	0
1984	15145	0	0.8	0.2	0	0	0	0
1985	15163	0	0.7	0.2	0	0	0	0
1986	15984	0	0	0	0.9	0	0	0
1987	16859	0	0	0	0	1	0	0
1988	18150	0	0	0	0	0	1	0
1989	18970	0	0	0	0	0	0	1
1990	19328	0	0	0	0	0	0	0.9
1991	19337	0	0	0	0	0	0	0.9
1992	18876	0	0	0	0	0	0.1	0.9

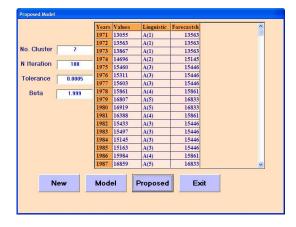


Figure 1. Forecasting enrollments of the University of Alabama by the proposed model

TABLE II. DATA ENROLLMENTS THE UNIVERSITY OF ALABAMA, LINGUISTIC VALUES, AND FORECASTED VALUES

Years	Enrollments	Linguistic	Forecasted
1971	13055	A_1	13563
1972	13563	A_1	13563
1973	13867	A_1	13563
1974	14696	A_2	15145
1975	15460	A_3	15446
1976	15311	A_3	15446
1977	15603	A_3	15446
1978	15861	A_4	15861
1979	16807	A_5	16833
1980	16919	A_5	16833
1981	16388	A_4	15861
1982	15433	A_3	15446
1983	15497	A_3	15446
1984	15145	A_3	15446
1985	15163	A_3	15446
1986	15984	A_4	15861
1987	16859	A_5	16833
1988	18150	A_6	18150
1989	18970	\mathbf{A}_7	18970
1990	19328	A_7	18970
1991	19337	A_7	18970
1992	18876	\mathbf{A}_7	18970

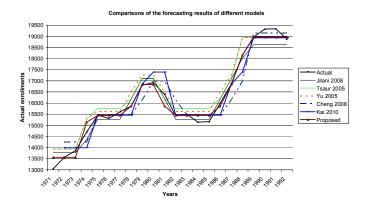


Figure 2. Forecasting results curve of enrollments of the university of Alabama

The forecasting value for year 1971 is 13563 while the actual value was 13055. Fig.1 and Table II show linguistic terms and forecasting values deduced by proposed model.

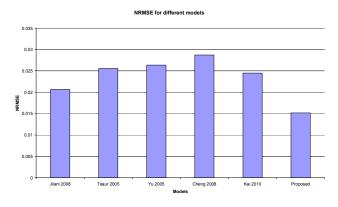


Figure 3. NRMSE-chart for the existing models and the proposed model

also shows that, the proposed model can further improve the forecasting results than the other model.

Fig. 3 shows the comparisons among the existing models by using NRMSE, where Jilani and Burney [7] model has 0.02, Tsaur and Yang [14] model has 0.025, Yu [15] model has 0.026, Kai et al [8] model has 0.024, Cheng, et al [4] model has 0.028 and proposed model has 0.015.

The line-chart comparison in Fig. 2 shows that the proposed model has higher accuracy than the other models. And the empirical comparison among the existing models in Table III

TABLE III. FORECASTING ENROLLMENTS OF THE UNIVERSITY OF ALABAMA

Year	Actual enrollments	Tsaur and Yang (2005)	Yu (2005)	Jilani and Burney (2008)	Cheng et al (2008)	Kai et al (2010)	Proposed
1971	13055	13934	13934	13769			13563
1972	13563	13934	13934	13769	14242	13997	13563
1973	13867	13934	13934	13769	14242	13997	13563
1974	14696	15298	15298	14360	14242	13997	15145
1975	15460	15753	15623	15271	15474.3	15461.2	15446
1976	15311	15753	15623	15271	15474.3	15461.2	15446
1977	15603	15753	15623	15271	15474.3	15461.2	15446
1978	15861	16208	16511	16182	15474.3	15461.2	15861
1979	16807	17118	17269	17094	16146.5	16861.7	16833
1980	16919	17118	17269	17094	16988.3	17394	16833
1981	16388	16208	16511	16182	16988.3	17394	15861
1982	15433	15753	15623	15271	16146.5	15461	15446
1983	15497	15753	15623	15271	15474.3	15461.2	15446
1984	15145	15753	15623	15271	15474.3	15461.2	15446
1985	15163	15753	15623	15271	15474.3	15461.5	15446
1986	15984	16208	16511	16182	15474.3	15461.5	15861
1987	16859	17118	17269	17094	16146.5	16861.7	16833
1988	18150	18937	18937	18004	16988.3	17394	18150
1989	18970	18937	18937	18624	19144	18932.2	18970
1990	19328	18937	18937	18624	19144	18932.2	18970
1991	19337	18937	18937	18624	19144	18932.2	18970
1992	18876	18937	18937	18624	19144	18932.2	18970
1	NRMSE	0.025	0.026	0.02	0.028	0.024	0.015

IV. EMPIRICAL STUDY

Based on the data of the iron and steel production witch are provided by the International Iron and Steel Institute in Brussels, Belgium, and publications of the U.S. geological survey from 1975 to 2008 (production values in thousand metric tons), we can get the universe of discourse U=[457000, 954000], partition U into 7 equal intervals, D_1 =6000, and D_2 =7000. Hence, the intervals are u_1 ; u_2 ; u_3 ; u_4 ; u_5 ; u_6 ; u_7 ; where :-

u_1 =[451000.00, 523857.14]
u ₂ =[523857.14, 596714.29],
u ₃ =[596714.29, 669571.43],
u ₄ =[669571.43, 742428.57],
$u_5 = [742428.57, 815285.71],$
$u_6 = [815285.71, 888142.86],$
$u_7 = [888142.86, 961000.00],$

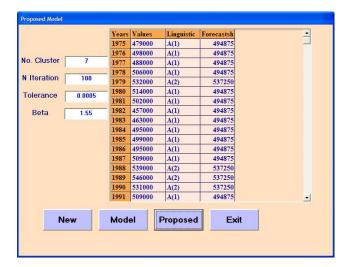


Figure 4. Forecasting of the world production of iron and steel by the proposed model

Table IV lists the World Production of Iron and Steel from 1975 to 2008, and membership grades of enrollments for each linguistic. Define the fuzzy set A_i using the linguistic variable "World Production of Iron and Steel", let A_1 = (very very few), A_2 = (very few), A_3 = (few), A_4 = (moderate), A_5 = (many), A_6 = (many many), A_7 = (too many).

Fig. 4 and Table V show linguistic terms and forecasting values deduced by proposed model. The forecasting value for year 1975 is 494875 while the actual value was 479000 and the forecasting value for year 2008 is 943000 while the actual value was 932000.

TABLE IV. DATA OF THE WORLD PRODUCTION OF IRON AND STEEL, AND MEMBERSHIP GRADES.

	MEMBERSHIP GRADES.									
Year	Production	$\mathbf{A_1}$	\mathbf{A}_{2}	\mathbf{A}_3	A_4	\mathbf{A}_{5}	A_6	A ₇		
1975	479000	1	0	0	0	0	0	0		
1976	498000	1	0	0	0	0	0	0		
1977	488000	1	0	0	0	0	0	0		
1978	506000	0	0	0	0	0	0	0		
1979	532000	0	1	0	0	0	0	0		
1980	514000	0	0	0	0	0	0	0		
1981	502000	1	0	0	0	0	0	0		
1982	457000	0	0	0	0	0	0	0		
1983	463000	0	0	0	0	0	0	0		
1984	495000	1	0	0	0	0	0	0		
1985	499000	1	0	0	0	0	0	0		
1986	495000	1	0	0	0	0	0	0		
1987	509000	0	0	0	0	0	0	0		
1988	539000	0	1	0	0	0	0	0		
1989	546000	0	1	0	0	0	0	0		
1990	531000	0	1	0	0	0	0	0		
1991	509000	0	0	0	0	0	0	0		
1992	503000	1	0	0	0	0	0	0		
1993	507000	0	0	0	0	0	0	0		
1994	516000	0	0	0	0	0	0	0		
1995	536000	0	1	0	0	0	0	0		
1996	516000	0	0	0	0	0	0	0		
1997	540000	0	1	0	0	0	0	0		
1998	535000	0	1	0	0	0	0	0		
1999	539000	0	1	0	0	0	0	0		
2000	573000	0	0	1	0	0	0	0		
2001	585000	0	0	1	0	0	0	0		
2002	608000	0	0	1	0	0	0	0		
2003	673000	0	0	0	1	0	0	0		
2004	720000	0	0	0	1	0	0	0		
2005	802000	0	0	0	0	1	0	0		
2006	881000	0	0	0	0	0	1	0		
2007	954000	0	0	0	0	0	0	1		
2008	932000	0	0	0	0	0	0	1		

The proposed model selected the maximum membership grade for each cluster, the forecasting value for each cluster calculating as in (7):

$$forecaste(A_1) = \frac{1 \times (1975) + 1 \times (1976) + 1 \times (1977) + 1 \times (1981) + 1 \times (1984) + 1 \times (1985) + 1 \times (1986) + 1 \times (1992)}{8} = 494875$$

$$forecaste(A_2) = \frac{1 \times (1979) + 1 \times (1988) + 1 \times (1989) + 1 \times (1990) + 1 \times (1995) + 1 \times (1997) + 1 \times (1998) + 1 \times (1999)}{8} = 537250$$

$$forecaste(A_3) = \frac{1 \times (2000) + 1 \times (2001) + 1 \times (2002)}{3} = 588667$$

$$forecaste(A_4) = \frac{1 \times (2003) + 1 \times (2004)}{2} = 696500$$

$$forecaste(A_5) = \frac{1 \times (2005)}{1} = 802000$$

$$forecaste(A_6) = \frac{1 \times (2006)}{1} = 881000$$

$$forecaste(A_7) = \frac{1 \times (2007) + 1 \times (2008)}{2} = 943000$$

TABLE V. DATA OF THE WORLD PRODUCTION OF IRON AND STEEL, LINGUISTIC VALUES, AND FORECASTED VALUES

Year	Production	Linguistic	Forecasted
1975	479000	\mathbf{A}_1	494875
1976	498000	\mathbf{A}_1	494875
1977	488000	\mathbf{A}_1	494875
1978	506000	\mathbf{A}_1	494875
1979	532000	A_2	537250
1980	514000	\mathbf{A}_1	494875
1981	502000	\mathbf{A}_1	494875
1982	457000	\mathbf{A}_1	494875
1983	463000	\mathbf{A}_1	494875
1984	495000	\mathbf{A}_1	494875
1985	499000	\mathbf{A}_1	494875
1986	495000	\mathbf{A}_1	494875
1987	509000	\mathbf{A}_1	494875
1988	539000	A_2	537250
1989	546000	A_2	537250
1990	531000	A_2	537250
1991	509000	\mathbf{A}_1	494875
1992	503000	\mathbf{A}_1	494875
1993	507000	\mathbf{A}_1	494875
1994	516000	\mathbf{A}_1	494875
1995	536000	A_2	537250
1996	516000	\mathbf{A}_{1}	494875
1997	540000	A_2	537250
1998	535000	\mathbf{A}_2	537250
1999	539000	A_2	537250
2000	573000	A_2	537250
2001	585000	\mathbf{A}_2	537250
2002	608000	A_3	588667
2003	673000	A_4	696500
2004	720000	A_4	696500
2005	802000	A_5	802000
2006	881000	A_6	881000
2007	954000	A_7	943000
2008	932000	A_7	943000

The researchers used famous models: Huarng[6], Tsaur and Yang [14], Yu [15], Jilani and Burney [7] to test the proposed model by forecasting of the world production of iron and steel as in Table VI.

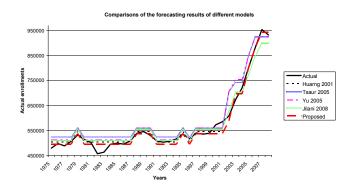


Figure 5. Forecasting results curve of the world production of iron and steel

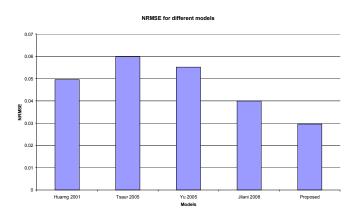


Figure 6. NRMSE-chart for the existing models and the proposed

The line-chart comparison in Fig. 5 shows that the proposed model has higher accuracy than the other models. And the empirical comparison among the existing models in Table VI also shows that, the proposed model can further improve the forecasting results than the other model.

Fig. 6 shows the comparisons among the existing models by using NRMSE, where Huarng[6] model has 0.0496, Tsaur and Yang [14] model has 0.0598, Yu [15] model has 0.0551, Jilani and Burney [7] model has 0.0399, and proposed model has 0.0296.

TABLE VI. FORECASTING OF THE WORLD PRODUCTION OF IRON AND STEEL

Year	Actual	Huarng 2001	Tsaur 2005	Yu 2005	Jilani 2008	Proposed
1975	479000	504571	523857	510762	509514	494875
1976	498000	504571	523857	510762	509514	494875
1977	488000	504571	523857	510762	509514	494875
1978	506000	504571	523857	510762	509514	494875
1979	532000	545714	560286	560286	555508	537250
1980	514000	504571	523857	510762	509514	494875
1981	502000	504571	523857	510762	509514	494875
1982	457000	504571	523857	510762	509514	494875
1983	463000	504571	523857	510762	509514	494875
1984	495000	504571	523857	510762	509514	494875
1985	499000	504571	523857	510762	509514	494875
1986	495000	504571	523857	510762	509514	494875
1987	509000	504571	523857	510762	509514	494875
1988	539000	545714	560286	560286	555508	537250
1989	546000	545714	560286	560286	555508	537250
1990	531000	545714	560286	560286	555508	537250
1991	509000	504571	523857	510762	509514	494875
1992	503000	504571	523857	510762	509514	494875
1993	507000	504571	523857	510762	509514	494875
1994	516000	504571	523857	510762	509514	494875
1995	536000	545714	560286	560286	555508	537250
1996	516000	504571	523857	510762	509514	494875
1997	540000	545714	560286	560286	555508	537250
1998	535000	545714	560286	560286	555508	537250
1999	539000	545714	560286	560286	555508	537250
2000	573000	545714	560286	560286	555508	537250
2001	585000	545714	560286	560286	555508	537250
2002	608000	706000	706000	706000	628923	588667
2003	673000	742429	742429	754571	702221	696500
2004	720000	742429	742429	754571	702221	696500
2005	802000	851714	851714	851714	775435	802000
2006	881000	924571	924571	924571	848587	881000
2007	954000	924571	924571	924571	898939	943000
2008	932000	924571	924571	924571	898939	943000
NR	MSE	0.0496	0.0598	0.0551	0.0399	0.0296

V. DISCUSSION AND CONCLUSION

The research proposed an efficient fuzzy time series forecasting model based on fuzzy clustering with high accuracy. The method of FCMI is integrated in the processes of fuzzy time series to partition datasets. Experimental results of enrollments of the University of Alabama, and the comparison between the existing models: Jilani and Burney [7], Tsaur and Yang [14], Yu [15], Kai et al [8], and Cheng, et al [4] and the proposed model show that, the proposed model can further improve the forecasting results than the other models and also the experimental results of the world production of iron and steel, and the comparison between the existing models: Huarng[6], Tsaur and Yang [14], Yu [15], Jilani and Burney[7] and the proposed model show that, the proposed model has higher accuracy than the other models.

VI. REFERENCES

- A. K. Abd Elaal, H. A. Hefny, and A. H. Abd-Elwahab, "A novel forecasting fuzzy time series model", in: Proceeding of International Conference on Mathematics and Information Security, Sohag Univ., Egypt, 2009.
- [2] A. K. Abd Elaal, H.A. Hefny, and A. H. Abd-Elwahab, "Constructing Fuzzy Time Series Model Based on Fuzzy Clustering for a Forecasting", J. Computer Sci., vol. 7, 2010, pp. 735-739.
- [3] T.-L. Chen, C.-H. Cheng, and H.-J. Teoh, "High-order fuzzy timeseries based on multi-period adaptation model for forecasting stock markets", Physica A, vol.387, 2008, pp. 876–888
- [4] C.-H. Cheng, J.-W. Wang, and G.-W. Cheng, "Multi-attribute fuzzy time series method based on fuzzy clustering", Expert Systems with Applications, Vol.34, 2008. pp. 1235–1242.
- [5] M. Friedman and A. Kandel, "Introduction to pattern recognition statistical, structural, neural and fuzzy logic approaches", Imperial college press, London, 1999, p. 329.
- [6] K. Huarng, "Effective lengths of intervals to improve forecasting in fuzzy time series", Fuzzy Sets and Systems, vol.123, 2001, pp. 387– 394
- [7] T.A. Jilani and S. Burney, "Multivariate stochastic fuzzy forecasting models", Expert Systems with Applications, vol.35, 2008, pp. 691– 700
- [8] Kai, F. Fang-Ping, and C. Wen-Gang, "A novel forecasting model of fuzzy time series based on K-means clustering", IWETCS, IEEE, 2010, pp.223–225.
- [9] G. Kirchgässner and J. Wolters, "Introduction to modern time series analysis", Springer-Verlag, Berlin, Germany, 2007, p.153.

- [10] H.-T. Liu, "An improved fuzzy time series forecasting method using trapezoidal fuzzy numbers". Fuzzy Optimization and Decision Making, vol. 6, 2007, pp.63-80.
- [11] A.K. Palit and D. Popovic, "Computational intelligence in time series forecasting theory and engineering applications", Springer-Verlag.London, UK, 2005, p.18.
- [12] Q. Song and B.S. Chissom, "Forecasting enrollments with fuzzy time series. I", Fuzzy sets and systems, vol. 54, 1993, pp. 1-9.
- [13] Q. Song and B.S. Chissom, "New models for forecasting enrollments: fuzzy time series and neural network approaches", ERIC, 1993 p. 27, http://www.eric.ed.gov
- [14] R.-C. Tsaur, J.-C. Yang, and H.-F. Wang, "Fuzzy relation analysis in fuzzy time series model", Computers and Mathematics with Applications, vol.49, 2005, pp. 539-548.
- [15] H.-K. Yu, "Weighted fuzzy time series models for TAIEX forecasting", Physica A, vol.349, 2005, pp.609–624.

AUTHORS PROFILE



Mr. Ashraf Khalaf Abd Elaal is a Ph.D. student in Computer Sciences Department at the Institute of Statistical Studies and Research, Cairo University. His Ph.D. in the filed of Computational Intelligence. His research interests include fuzzy time series, Fuzzy clustering



Dr. Hesham Ahmed Hefny is an assistant professor and the head of Computer & Information Sciences Department at the Institute of Statistical Studies and research, Cairo University. His research interests include Artificial Neural Networks, Fuzzy Systems, Genetic Algorithms, Swarm Intelligence, Pattern Recognition, and Data Mining. Dr.

Hesham has published over 35 peer refereed papers in academic journals and conferences on topics within Artificial Intelligence and related areas.



Prof. Ashraf Hassan Abdelwahab is a professor of computer engineering, Electronics Research Institute, Cairo, Egypt. He received his M. Sc. in 1988, Faculty of Engineering, Cairo University in the area of Artificial Intelligence, and in 1992 he received his Ph.D. in Machine Learning and Evolutionary Algorithms. He has published

over 60 technical papers in National, and International journals and conferences in the areas of Evolutionary Algorithms, Machine Learning, and Data Mining.

The 2D Image-Based Anthropologic Measurement by Using Chinese Medical Acupuncture and Human Body Slice Model

Sheng-Fuu Lin
Institute of Electrical Control Engineering
National Chiao Tung University
Hsinchu, Taiwan(ROC)

Shih-Che Chien Institute of Electrical Control Engineering National Chiao Tung University Hsinchu, Taiwan(ROC)

Kuo-Yu Chiu Institute of Electrical Control Engineering National Chiao Tung University Hsinchu, Taiwan(ROC)

Abstract-Human body anthropometric measurement is widely used in daily life and has become an indispensable item for people and garment manufactures. Two-dimensional image-based anthropometric measurement system provides another choice to make anthropometric measurement alternative to traditional and three-dimension methods. These measurement systems are attractive because of their lower cost and easier to use. Although these systems have appeared in this type of application, most systems require the user to wear as little as possible for reducing the errors which come from garments in the measurement, and furthermore the measurement equipments are not easy available at anywhere and the setup are always complex. This paper presents an approach with fewer constraints and more simplified operation, and has the performance as good as manual measurement. In this approach, the Chinese medicine acupuncture theory is used to locate the position which measurement concerned and replace the manual marking or other feature extraction methods. For circumferences measurement, the human body slice model is supplied to approach the circumference shapes and used the piecewise Bezier curve to approximate the circumference curve. At the final, a compensation system of garment thickness is employed to amend the measurement data, which are obtained through the direct measuring of subject wearing clothes, to ensure the accuracy. Through of these methods, the subjects for measurement are not required for wearing as little as possible and the results of experiments also shows that the approach is quite comparable to traditional measurement methods.

Keywords- anthropometric measurement; Chinese medicine acupuncture; garment thickness

I. INTRODUCTION

A For long time ago, people knew how to use tools to take a measurement of human body, stature, lengths, and circumferences, and use these anthropometric data for manufactory product, garment designed, statistic...etc. Nowadays, the anthropologic measurement is more widely used than before and become an indispensable item in human life. With the rapid development of internet worldwide, the consumption pattern has been changed. The virtual shop online is a novel type of shopping platform that the customer can make a purchase at anytime, anywhere, and no distance limit thought internet network. Hence, the method for anthropometric measurement without complex setting and no experience required is necessary.

In traditional anthropometric measurement methods, the measurement can be done by using a simple instrument such as tape and without complex measurement pre-setting. The anthropometric data can be classified into two types: linear distance and circumference. The linear part is defined as the distance between two anatomical landmarks. The circumference part is defined as the length that can close around the part at predefined location. Through the use of manual measurements by domain experts, combined with precise measurement devices, an accurate model of human body and its measurement data can be obtained [1]. Although the traditional anthropometry measurement is easy and convenient to use, the traditional measurement relies on manual operations that are inefficient and prone to errors.

2D image-based anthropometric measurement methods adopt two or more photographic to do the anthropometric measurement. Generally, the camera must be set and calibrated before capturing image. Through image process and geometric transformation, the anthropometric measurement data can be obtained from captured images. Meunier and Yin [2] proposed an anthropometric measurement system that can generate body measurements from two-dimension images. Their system comprised two synchronized and calibrated cameras that captured two images: front and side. Six anthropometric measurement data (neck circumference, chest circumference, hip circumference, waist circumference, stature, and sleeve length) can be measured and compared with manual measurement by using tape for efficient evaluation. Hung et al. [3] used geometric shapes (ellipse and rectangular) to approximate the shape of critical part circumference. The ellipse shape was used to approximate neck, wrist, and palm circumferences, and the major and minor axes length were obtained from the front and side views. The combination of a rectangle and an ellipse was used to approximate the chest circumference, and with the calculation of perimeter the chest circumference can be obtained. Although, 2D image-based methods can be used for anthropometric measurement, the complicated pre-setting, camera setting and calibration, and too much manual operations make it not easy for people to use.

There have been significant hardware and software development in laser scanning technique during the last two decades. Nowadays, the 3-D laser scanning technique makes it possible to digitalize the complete surface of human body and provides much richer information about the body shape than the traditional anthropometric measurements [4]. CAESAR, which stands for Civilian American and European Surface Anthropometry Resource, is the first large scale 3-D anthropometry survey project [5]. This approach provides a standard model of digitalized human body shape and opens up opportunities to extract new measurements for quantifying the body shape. The first attempt in processing 3-D anthropometric data for analyzing the body shape is to extract traditional anthropometric measurements from the scanned data [6]. Although working with the 3-D surface data has the advantage of being able to perform repeated measurements without the subject being present, the 3-D scanning equipment is not available in anytime and anywhere and the collected 3-D scanning data still needs complicated analysis to calculate human body anthropometric data. Furthermore, the 3-D human body scanning data would possibly touch off privacy right because of its rich and detail information of human body shape.

This paper describes an approach to make an anthropometric measurement with fewer constraints, no high-price measurement instruments and complicated pre-setting. The Chinese medicine acupuncture theory, which has a distinctive way to locate the acupuncture points on human body, is used to fast locate the position of critical parts. The human body slice model accessed from the slice shape analysis of 3-D scanning data at critical part is applied to approximate the circumference shape. The fuzzy inference system is adopted to evaluate the influence of garment for anthropometric measurement, and the compensation system is applied to correct the measured data. The objective of this paper is to compare the accuracy and precision of measurements made from 2D images of human with those of highly trained anthropometric expert [2].

This paper is organized as follows. Section II provides a 2D image-based anthropometric measurement that used two side images, front and side, which captured by camera, and the human body slice model is supplied to approximate the critical circumference shape. In section III, the compensation system, including garment thickness estimation and measurement compensation, is proposed. The experimental results and discussion are showed in section IV and section V respectively. Section VI concludes this paper.

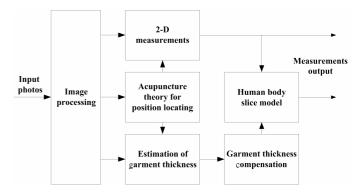


Figure 1. The flowchart of 2-D image-based anthropometric measurement.

II. 2D IMAGE-BASED ANTHROPOMETRIC MEASUREMENT

In this section, the method of 2D image based anthropometric measurement is provided, and includes system description. image preprocessing, Chinese acupuncture theory for position locating, and human body slice model. At the first, system description is to descript the measurement instruments and the requirements of system. Secondly, image preprocessing is used to process the images captured from camera and acquire the information to later analysis. Thirdly, the Chinese medicine acupuncture theory is adopted to locate the critical position because of its distinctive acupuncture point locating method of human body. After locating the critical position, the 2D anthropometric data can be obtained by direct measurement. In order to acquire the circumference, the human body slice model is used to approximate the circumference shape. The flowchart of this system illustrated in Fig. 1.

A. System description

The system under review is a PC-based system and comprised of one Fuji F40 color digital camera (1280 x 960 pixels) and a white backdrop. The camera is set on tripod with about one meter off the ground, and dual-axis bubble levels on tripod help to keep the camera level. The system captures front and side images of objects standing with some required postures. In the front view, the object stand with their arms straight and abducted about 45 degree angles from body, and the legs straight and slightly abducted. In the side view, the object stand with their arms straight and close to body, and the legs straight and slightly abducted. In both views, the object is required for all fingers closing together without bending, eyes looking forward, no glove, and no shoes. The requirement of palm is in order to measure the width of four fingers (forefinger, middle finger, ring finger, and little finger), which can be used for computing the spatial resolution and locating position and will be described in next paragraph and Section II.C. Fig. 2 shows the required postures of object in two views.

In this paper, the system is assumed to know the anthropometric data (stature, weight, and width of four fingers) of individual at the beginning. Through the known anthropometric data, the Body Mass Index (BMI) can be calculated and applied for rough classification of human body.





Figure 2. The postures of front and side view.

The spatial resolution method, the ratio for direct mapping of pixel to actual length, is used to replace the camera calibration for simplification. By that way, the actual length of object can be calculated if the pixels belong to object length in the image is knew. Since the actual width of four fingers is known, the spatial resolution of system can be obtained by counting the pixels belong to width of four fingers in image.

B. System description

The purposes of this part are to extract the human body shape from captured image and label the skin regions, where without clothes covered. To extract the correct human body shape is very important for 2D anthropometric measurement, because the error at the extracted human body shape will be amplified in final result of anthropometric data. The extraction of human body shape is performed by automatic image segmentation [7][8], which uses color edge extraction and seeded region growing (SRG) to do image segmentation. The purpose of SRG is to merge the part of the same homogeneous region. It is also possibility that some regions are actually part of the same homogeneous region but have been split because the initial given seeds for them are different. Hence, the selected region seed should represent the distinguishing character of the corresponding region. For our application, the centroids of head, limbs, and body trunk are good as the seeds for object generation. So we chose six centriods of head, right and left hand (palm), body trunk, and right and left foot (sole of foot) as seeds for region growing. Except for the centriod of body trunk, the regions for seed selected are all belong to skin region. For this reason, the skin detection method developed by Albiol et al. [8] is used to extract skin color pixels and median filter is applied to eliminate the isolated pixel. Through the connected component labeling method, the skin region can be labeled as head, right and left hand (palm), and right and left foot (sole of foot). The centriod of body trunk is defined as the mass center of triangle which vertices are the centriods of head, right hand, and left hand. Fig. 3(a) shows the edges of captured front image by performing the Sobel edge detector and Fig. 3(b) shows the obtained skin area. By performing the SRG, the human body mask, Fig. 3(c), can be acquired.

In order to access the spatial resolution described in section II.A, the pixel belong to four fingers width must be counted. The definition of four-finger width is the width of four closed fingers at the position of second forefinger knuckle and Fig. 4 illustrates the four-finger width. Actually, the correct position

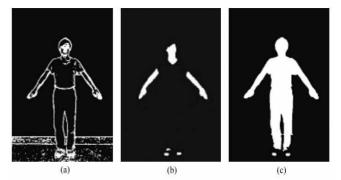


Figure 3. The segmentation of the body from the background. (a) edges (b) skin area (c) human body mask

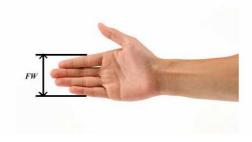


Figure 4. Illustrate the definition of four-finger width.

of second knuckle of forefinger is hard to recognize, especially in silhouette. Hence, the four-finger width is calculated for the average width at the restricted region around second knuckle of forefinger.

There are two steps to extract the width of four fingers in 2D image. The first step is to rotate the right hand skin region to horizontal with the rotation angle, which is defined as the angle between the axis of right hand skin region and horizontal axis of image. The second step is to make a histogram of horizontal right hand skin region, and the average width for four-finger width *FWavg*, can be defined as:

$$FW_{avg} = \frac{1}{b - a + 1} \sum_{i=a}^{b} P(i)$$
 (1)

$$a = n + round(\frac{m-n}{3}), b = n + round(\frac{2(m-n)}{3})$$
 (2)

where P(i) denotes the histogram value at the position i, n is the position of fingertip of middle finger, and m is the position where has the maximum histogram value. Take Fig. 5(a) as an example, the measured data n and m are 40 and 72, and the data a and b can also be calculated as 51 and 61 respectively.

C. Chinese medicine acupuncture theory for locating position

Acupuncture is originated in China and has a long history. Along with the accumulation of clinical experience, acupuncture treatment has become a developed theory. Acupuncture is the procedure of inserting filiform needles into various acupuncture points on the body to relieve pain or for

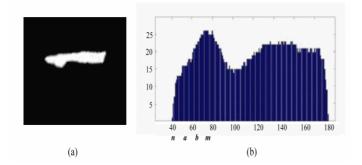


Figure 5. Illustrate the procedure of four-finger width calculation. (a) the horizontal right hand skin area and (b) histogram of (a)

therapeutic purposes. In treatment procedural, the therapist should find out the correct acupuncture point position corresponding to pain before inserting needle. Hence, the acupuncture point location method is developed to locate the acupuncture points on human body fast and accurately. The most popular used method of acupuncture point location is finger-inch [9], which uses a thumb width as the measurement unit and denotes the distance between acupuncture point and neighbor acupuncture points with measurement unit. In generally, the width of four fingers (forefinger, middle finger, ring finger, and little finger), has been described and defined in past paragraph, can be denoted as 3 units in finger-inch method.

In this paper, five body measurement data (shoulder length, chest circumference, waist circumference, hip circumference, and leg length) are selected for discussion because there are the most commonly used in manufactory and general application, such as clothing size. For locating the corresponding position, six acupuncture points (Lian-Quan, Tian-Tu, Chien-Yu, Shan-Zhong, Shen-Que, and Qu-Gu) are taken to locate the critical positions. The position of Lian-Quan acupuncture point is near the center of hyoid bone, so it can be represented as the point of chin. The position of Tian-Tu is very close to center of two Chien-Yu acupuncture points which are located at point of shoulder joints and the distance of these two acupuncture points is denoted as the shoulder length. The position of Shan-Zhong acupuncture point is at the center of nipples. Therefore, Shan-Zhong acupuncture point can be denoted as the position of chest. The positions of Shen-Que acupuncture point and navel are virtually identical, so it can be denoted as the position of waist. The position of Qu-Gu acupuncture point is near the upper edge of pubic symphysis, and could be defined as the position of hip. Because these five acupuncture points are all located at the trunk axis line of human body, the sequence of these four acupuncture from top to down (Lian-Quan, Tian-Tu, Shan-Zhong, Shen-Que, and Qu-Gu) could be obtained and the relative distance are two units, five units, eleven units, and five units of finger-inch respectively. The Fig. 6 shows the position of acupuncture point on human body.

D. 2D anthropometric measurement

The measurement data can be separated into two groups, direct measurement (linear) and indirect measurement (circumference). In linear part, the measurement is to calculate the distance between two acupuncture points: shoulder length is

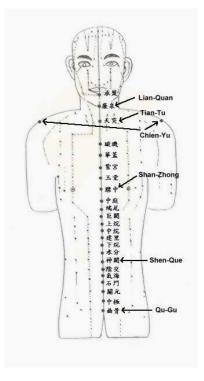


Figure 6. Illustrate the position of acupuncture points on human body.

the distance between two Chien-Yu acupuncture points or to calculate the distance between the acupuncture point and extremity: leg length is the distance between the Qu-Gu acupuncture point and sole of foot. Circumference cannot be measured directly using front and side view images only and must therefore be calculated using some form of mathematical model. In this paper, the human body slice model is adopted to approximate the circumference shape, and the circumference could be obtained through the calculation with two parameters, the corresponding width of front and side view image.

E. 2D anthropometric measurement

The measurement data can be separated into two groups, direct measurement (linear) and indirect measurement (circumference). In linear part, the measurement is to calculate the distance between two acupuncture points: shoulder length is the distance between two Chien-Yu acupuncture points or to calculate the distance between the acupuncture point and extremity: leg length is the distance between the Qu-Gu acupuncture point and sole of foot. Circumference cannot be measured directly using front and side view images only and must therefore be calculated using some form of mathematical model. In this paper, the human body slice model is adopted to approximate the circumference shape, and the circumference could be obtained through the calculation with two parameters, the corresponding width of front and side view image.

F. Human body slice model

Circumference cannot be measured directly using only 2D measurement, and therefore the mathematical model is applied to approximate the corresponding circumference. In this paper,

the human body slice model which using Bezier curve [10] is used to approximate the circumference. Bezier curve is wildly used in computer graphics to model smooth curves, and can be applied to approximate the curve by using the respective transform on the control points. The quadratic and cubic Bezier curve functions are most common used, because the higher degree Bezier curve function is more expensive to evaluate. If the more complex shape is needed, the lower degree Bezier curve can be patched together, which also called piecewise polynomial curve. In this paper, the cubic Bezier curve is adopted for basic curve to patch the piecewise curve. The cubic Bezier curve content with four control points: P_0 , P_1 , P_2 , and P_3 , which existing on the same plane of curve. The cubic Bezier curve B(t) starts from P_0 to P_1 and arrive at P_3 , and the function can be denoted as follow:

$$B(t) = (1-t)^{3} P_{0} + 3(1-t)^{2} t P_{1} + 3(1-t)t^{2} P_{2} + t^{3} P_{3} , t \in [0,1].$$

$$= [t^{3} t^{2} t 1] \begin{bmatrix} -1 & 3 & -3 & 1 \\ 3 & -6 & 3 & 0 \\ -3 & 3 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} P_{0} \\ P_{1} \\ P_{2} \\ P_{3} \end{bmatrix}$$

$$(3)$$

The 3-D human body scanning technique makes it possible to digitalize the complete surfaces of large number of human bodies, and provides much richer information about the body shape than the traditional and 2D anthropometric measurement. If the complete digitalized human body data is available, the slice data of human body can be extracted with the plane which is concerned. The Taiwan human body bank (TAIBBK) [11][12], which stands for Industrial Technology Research Institute, Tshing Hua University, and Chang Gung University, is a large scale 3-D anthropometric measurement project. About 1100 civilians, between the age of 19 and 65 in Taiwan, were scanned.

Because the build of fat man and thin man or male and female is much different, the slice models of human body must be set up to different build types respectively. For this reason, 100 people (50 males and 50 females) were selected in TAIBBK database, and the males and females were divided into five clusters on the basis of BMI (BMI<18.5, $18.5 \le BMI < 25$, $25 \le BMI < 30$, $30 \le BMI < 35$, and $BMI \ge 35$) respectively. Then, the slice human model, which contain with shapes of chest, waist, and hip, of each build types can be obtained. Fig. 7 shows the 3-D virtual mannequin and corresponding slice shape of standard male build ($18.5 \le BMI < 25$).

In generally, the shape of human body is virtually bilateral reflection symmetry with respect to the center mirror plane. Furthermore, the shapes of chest, waist, and hip are also symmetrically in slice human model. It also means that the circumference can be obtained, if the curve length of corresponding half-shape is known. Fig. 8(a) shows the whole and half chest slice shape of standard male build type, and confirms the symmetry of slice shape model. In this paper, the piecewise polynomial curve based on cubic Bezier curve is used to approximate the curve of half-shape, and then the curve length can be calculated by using definite integral. Fig. 8(b)

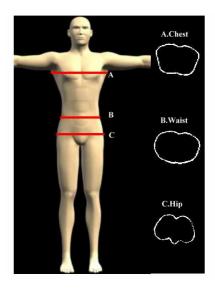


Figure 7. Illustrate the 3-D virtual mannequin, chest slice shape, waist slice shape, and hip slice shape of standard build of male.

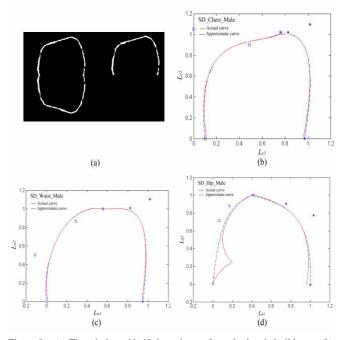


Figure 8. (a) The whole and half chest shape of standard male build type. (b) The approximation curve and actual curve of chest, (c) The approximation curve and actual curve of waist. (d) The approximation curve and actual curve of hip.

shows the curve approximation with two cubic Bezier curves, and two sets of control points are marked with blue circle and star. The red curve is actual curve of chest slice shape and the blue dotted curve is the approximation curve, which is similar to manual measurement.

Then, the expression of piecewise polynomial curve can be denoted as follow:

$$C(t) = C_{bz1} + C_{bz2} = TB_c (P_1 + P_2)$$
 (4)

where P1 and P2 are the control point sets of anterior curve (C_{bz1}) and posterior curve (C_{bz2}) respectively. Now, we take the chest circumference measurement of standard male build type for example. The expression of approximation curve, showed in Fig. 8(b) and followed from Eq. (4) is showed as:

$$C_{chest}(t) = TB_c (P_{c1} + P_{c2})$$
 (5)

$$P_{c1} = \begin{bmatrix} 0.1 & 0 \\ 0 & 1.05 \\ 0.490.88 \\ 0.75 & 1 \end{bmatrix} \begin{bmatrix} L_{c1} & 0 \\ 0 & L_{c2} \end{bmatrix}$$

$$P_{c2} = \begin{bmatrix} 0.75 & 1 \\ 0.82 & 1 \\ 1.141.11 \\ 0.07 & 0 \end{bmatrix} \begin{bmatrix} L_{c1} & 0 \\ 0 & L_{c2} \end{bmatrix}$$

$$(6)$$

$$P_{c2} = \begin{bmatrix} 0.75 & 1\\ 0.82 & 1\\ 1.141.11\\ 0.97 & 0 \end{bmatrix} \begin{bmatrix} L_{c1} & 0\\ 0 & L_{c2} \end{bmatrix}$$
 (7)

where L_{c1} and L_{c2} are the width of chest in side image and half width of chest in front image respectively. Similarly, the approximate curves of waist and hip of standard man build type are show in Fig. 8(c) and Fig. 8(d) respectively, and the approximate curve of waist is given by expression

$$C_{waist}(t) = TB_c(P_{w1} + P_{w2})$$
 (8)

$$P_{w1} = \begin{bmatrix} 0 & 0 \\ -0.11 & 0.5 \\ 0.29 & 0.88 \\ 0.58 & 1 \end{bmatrix} \begin{bmatrix} L_{w1} & 0 \\ 0 & L_{w2} \end{bmatrix}$$

$$P_{w2} = \begin{bmatrix} 0.58 & 1 \\ 0.82 & 1 \\ 1.071.11 \end{bmatrix} \begin{bmatrix} L_{w1} & 0 \\ 0 & L_{w2} \end{bmatrix}$$

$$(9)$$

$$P_{w2} = \begin{bmatrix} 0.58 & 1\\ 0.82 & 1\\ 1.071.11\\ 0.95 & 0 \end{bmatrix} \begin{bmatrix} L_{w1} & 0\\ 0 & L_{w2} \end{bmatrix}$$
 (10)

where L_{w1} and L_{w2} are the width of waist in side image and half width of waist in front image respectively, and the approximate curve of hip is given by expression

$$C_{hip}(t) = TB_c(P_{h1} + P_{h2}) \tag{11}$$

$$P_{h1} = \begin{bmatrix} 0 & 0 \\ 0.680.75 \\ 0.180.88 \\ 0.4 & 1 \end{bmatrix} \begin{bmatrix} L_{h1} & 0 \\ 0 & L_{h2} \end{bmatrix}$$
 (12)

$$P_{h2} = \begin{bmatrix} 0.4 & 1 \\ 0.75 & 0.9 \\ 1.03 & 0.79 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} L_{h1} & 0 \\ 0 & L_{h2} \end{bmatrix}$$
 (13)

where Lh1 and Lh2 are the width of hip in side image and half width of hip in front image respectively.

III. THE COMPENSATION SYSTEM FOR MEASUREMENT

In general 2D image-based anthropometric measurement, the subject has been required for no-wearing clothes or only wearing light underwear. Therefore, the application of

anthropometric measurement and the place where measurement can be used are limited. In order to deal with this issue, an efficient compensation system that aims to reduce the influence of subject wearing clothes for measurement is presented in this paper. This compensation system is composed of two parts; estimation of garment thickness and compensation of measurement. Because the garment has a great influence on measurement of chest circumference and waist circumference obviously, the compensations of chest and waist measurement are concerned here. The most of pants are close to buttock at hip, so the influence of pant on hip measurement is slight and can be neglected. The two direct measurement data, shoulder length and leg length, also could be neglected because of slight influence for distance measurement.

A. Estimation of garment thickness

Strictly, the influence of garment on chest and waist measurement does not only belong to garment thickness, but also to the properties of garment; such as the loose or fitting garment. But now we treat them as the same impact factor for garment measurement. Fuzzy inference systems have been successfully applied in fields of decision analysis and expert systems, and the Fuzzy inference is the process of formulating the mapping from a given input to an output using fuzzy logic. In this paper, the fuzzy inference system is used to infer the thickness of garment with given two inputs. The first input HC_{ratio} , is defined as the ratio of middle arm width to maximum value of possible width and expressed as:

$$HC_{ratio} = \begin{cases} 1 & , & \text{if } HC > 1.5FW_{avg} \\ \frac{HC - 0.6FW_{avg}}{HC_{max}}, & \text{otherwise} \end{cases}$$
(14)

where HC is the width of middle arm and HC_{max} is the maximum value of possible middle arm width. HC_{max} is given as 1.5 times of four-finger width here. The second input Diffratio is defined as the ratio of Diff, which is the difference in width between upper end of pant and lower end of garment, to maximum value of possible difference. The function of Diffratio is given by

$$Diff_{ratio} = \begin{cases} 1 & , & \text{if } Diff > 1.8FW_{avg} \\ \frac{Diff}{Diff_{max}}, & \text{otherwise} \end{cases}$$
(15)

where $Diff_{max}$ is the maximum value of possible Diff and given as 1.8 times of four-finger width.

Garment thickness (DT_{degree}) is the output of fuzzy inference system and lies in the interval 0 to 1, which represent the thinnest and thickest respectively. Follow the Eq. (14), the input value HC_{ratio} lies in the interval 0 to 1, which represent the few and many respectively. Similarly, the input value Diff_{ratio} lies in the interval 0 to 1, which represent the small and very large respectively. The membership functions of input and output are trapezoid type and triangle type respectively and show in Fig. 9. The fuzzy rules of this system are listed in TABLE I. Taking one rule for example, if HC_{ratio} is "few" and $Diff_{ratio}$ is "large", and then the DT_{degree} is "median". In defuzzification processes, the stander Mamdani -type method with COA-defuzzifier (center of area) is adopted.

TABLE I. THE FUZZY RULES

Diff _{ratio} HC _{ratio}	few	median	many
small	thiner	median	thicker
median	thin	thick	thicker
large	median	thicker	thickest
very large	thick	thickest	thickest

B. Compensation of measurement

The compensation system is a negative compensation and modification mechanism. The goal is to retrieval the chest and waist measured data in 2D measurement. So the outputs of compensation system are the reduction rates of measurement data, which are the direct measurement width of chest and waist in front and side view. The impact factors, also called inputs, of this system are garment thickness DT_{degree} and BMI.

The neural network is an efficiently method to simulate the structure and functional aspects of biological neural network. Hence, the back-propagation neural network (BPNN) [13] is used to structure compensation system. The tan-sigmoid function is applied for transfer function and 105 (63 males and 42 females) training data are used.

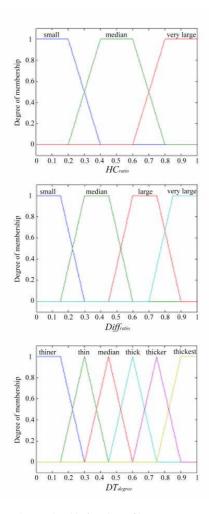


Figure 9. The membership functions of inputs HC_{ratio} and $Diff_{ratio}$ and output DT_{degree}

IV. EXPERIMENTAL RESULTS

This section describes the experiment results including accuracy and precision parts. The accuracy part adopts statistic methods to calculate the correlations between proposed system and traditional measurement system. The precision analysis is designed to repeated measurements by proposed and manual method, and aimed to show the repeatability of system.

A. Accuracy

The accuracy of the 2D image-based measurement system was calculated by comparing 2D image-based measurement with the manual measurement, or called traditional method, taken by anthropometric [14]. Two experiments, simplex garment, short sleeve T-shirt, and manifold garment, are performed in accuracy part. The first experiment is to calculate the accuracy of measurement system with less influence of garment, and the second experiment is to calculate the accuracy of measurement system with each kind of garment.

In the first experiment, the test sample composed of 175 subjects (97 males and 78 females) that have worn short sleeve T-shirt and been measured both with traditional method and with 2D image-based method. Each subject was measured once by each method. The means and standard deviations for subjects measured in these two methods are listed in TABLE II. The Pearson correlation coefficients between manual and 2D image-based measurements also list in TABLE II. The error rate (*E*) is the absolute value of difference of measurement and actual value (manual measurement) against the actual value. It is computed using the follow equation:

$$E = \frac{\left|x_{measurement} - x_{actual}\right|}{x_{actual}} \times 100\% . \tag{16}$$

In the second experiment, the test sample compose with 31 males and 15 females and be required to wear 4 kinds of garment, short T-shirt, long T-shirt, thin jacket, and thick jacket, for measurement. In order to assess the accuracies of 2D image-based measurement for each type of garment, the

TABLE II. ACCURACY RESULTS OF SIMPLEX GARMENT (SHORT SLEEVE T-SHIRT) (MM)

Measurement	Males	(n=97)			Female ((n=78)		
	Mean	SD	Corr.	E_{avg}	^S Mean	SD	Corr.	E_{avg}
Shoulder length								
Manual	443	21			407	22		
2-D system	442	19	0.90	2.17%	404	21	0.95	1.99%
Chest circ.								
Manual	1002	73			947	83		
2-D system	1006	70	0.96	1.78%	949	80	0.96	1.87%
Waist circ.								
Manual	999	79			935	71		
2-D system	1003	77	0.97	1.74%	937	70	0.98	1.71%
Hip circ.								
Manual	1011	72			974	79		
2-D system	1013	69	0.95	1.82%	973	77	0.96	1.83%
Leg length								
Manual	1018	28			977	26		
2-D system	1017	26	0.93	1.87%	979	24	0.92	1.88%

TABLE III. ACCURACY RESULTS OF MANIFOLD GARMENT (MM)

Measurement	(n=46)			
	Mean	SD	Corr.	E_{avg}
Shoulder length				
Manual	442	22		
2-D system I	440	20	0.91	1.92%
2-D system II	439	21	0.92	1.93%
2-D system III	440	21	0.91	2.11%
2-D system IV	445	22	0.90	2.29%
Chest circumference				
Manual	987	77		
2-D system I	990	76	0.96	1.77%
2-D system II	992	77	0.95	1.81%
2-D system III	992	76	0.95	2.37%
2-D system IV	995	81	0.90	4.43%
Waist circumference				
Manual	989	80		
2-D system I	992	80	0.98	1.66%
2-D system II	991	79	0.96	1.78%
2-D system III	995	81	0.97	3.29%
2-D system IV	996	86	0.88	4.58%
Hip circumference				
Manual	1003	76		
2-D system I	1005	77	0.96	1.88%
2-D system II	1004	77	0.95	1.93%
2-D system III	1005	75	0.92	1.77%
2-D system IV	1004	75	0.93	2.09%
Leg length				
Manual	1004	30		
2-D system I	1001	29	0.96	1.92%
2-D system II	999	28	0.95	1.97%
2-D system III	1001	30	0.97	2.89%
2-D system IV	1002	31	0.96	2.13%

(I: short sleeve T-shirt, II: long sleeve T-shirt, III: thin jacket, and IV: thick jacket)

system measurement and once with manual measurement. The results of means, standard deviations, average error rate, and Pearson correlation coefficients are listed in TABLE III.

B. Precision

The precision of the 2D image-based system was determined by performing ten repeat measurements on subject wore short sleeve T-shirt for manual and 2D image-based method. The camera setting and lighting conditions were relative constant and the re-measurement must be done within minute to ensure the postural of subject is almost the same. The means, standard deviations, and difference range of repeated measurement is listed in TABLE IV.

V. DISCUSSION

This section will discuss the experiment results, which contains accuracy, precision, and reliability. In the accuracy part, the statistical methods, t-test, f-test, and Pearson correlation coefficient, and simple error rate are adopted for assessment. The higher correlation coefficient and lower error rate indicate the higher accuracy. The precision is defined as the difference in values obtained from measuring the same

object repeatedly. The confidence interval method with 95% confidence level is applied to evaluate the precision of repeated measurement. In the same confidence level, the smaller confidence interval means the higher precision, or repeatability. The reliability coefficient is applied to measure the correlation of the repeated measurements and two useful parameters: the Technical Error of Measurement (TEM) and reliability coefficient (R) [15], are used to characterize reliability of system here.

TABLE IV. MEASUREMENT REPEATABILITY RESULTS (MM)

Measurement	Mean	Range	Std.	1.96 Std.
	(n=10)		Dev.	Dev.
Shoulder length				
Manual	438	3	1.87	3.67
2-D system	433	5	1.97	3.86
Chest circumference				
Manual	991	12	4.56	8.93
2-D system	1001	15	4.77	9.35
Waist circumference				
Manual	924	9	2.95	5.78
2-D system	938	12	3.46	6.78
Hip circumference				
Manual	983	8	2.06	4.03
2-D system	996	12	2.51	4.92
Leg length				
Manual	1004	9	2.26	4.43
2-D system	1017	11	2.45	4.80

A. Accuracy

Two statistical tests: *t*-test and *f*-test, and error rate are sufficient to characterize the accuracy of an anthropometric measurement. The *t*-test is performed to compare the means of all dimensions, and *f*-test is performed compare the standard deviation of all dimensions. The error rate is a simple way to quantify the accuracy and it is also easy to interpret the amount of difference between two measurements. The Pearson correlation coefficient is a measure of the strength of the association between the two set of measurement, and could also to be used to represent the degree of accuracy.

In first experiment results, *t*-tests and *f*-tests comparing of manual and 2D image-based measurements show no significant difference, indicating that the results of these two methods are similar and equally consistent. The results of Pearson correlation coefficient show that the closely perfect positive correlations were calculated and exhibited these two measurement data have similar. The results of average error rate were less than 2.5 %, and showed the perfect accuracy of 2D image-based measurement.

The second experiment performed to compare the 2D image-based measurement for each type of garment. Although the challenge of four type of garment was added in this experiment, the t-tests and f-tests comparing of manual and 2D image-based measurements show no significant difference (p<0.05), indicating that the similarity of results of these two methods is more than 95 %. In other words, the accuracy of 2D image-based measurement is more than 95%. The results of

Pearson correlation coefficient exhibit the highly positive correlation of 2D image-based measurement and manual measurement. No matter what type of garment was worn, the average error rates of each dimension are less than 5 %. It can be proved that the compensation system is feasible to reduce the influence of garment thickness on measurement.

Although the accuracy of 2D image-based system had been proofed to reach more than 95% from the results of two experiments, the error of measurement can not be ignored. The error in spatial resolution, which came from the calculation of four-finger width, will be spread to each measurement, and the effect of indirect measurement (circumference) is more than linear measurement in theoretical assessment. Within the same factor of error in spatial resolution, calculation of four-finger width, the error of critical position is built from the location procedure using Chinese medicine acupuncture theory.

The error of circumference measurement came from the imperfect approximation of circumference shape, and the integrities of inference system of garment thickness and compensation would cause the accuracy. In this paper, the human body slices model and inference systems are performed to indirect measurement and has good performance on accuracy, indicating that the adequate approximation model and inference systems are presented.

B. Precision

The precision is defined as the difference in values obtained from measuring the same object repeatedly. As show in TABLE IV, the results of repeated measurement showed the error of manual measurement (observed by skilled man) to be within around 3 to 9 mm of means, 95% of the time. For the most part, the linear measurements are more repeatable and precise than circumference, and the hip circumference are more repeatable and precise than other circumferences. The results in TABLE IV show that the repeated measurements by 2D imagebased method exhibited the error of precision to be within around 3 to 10 mm of means, 95% of the time, and showed the same basic trend as for the manual in that the linear measurements were more precise than circumferences, and hip circumference was more repeatable than other circumferences. The results of chest circumference and waist circumference in both methods exhibited more variability than others measurements, which was anticipated. The reason of this can be partly interpreted by torso movement due to the expansion and contraction of ribcage and abdomen of breathing and differences in posture of subject.

C. Reliability

The reliability is a kind of correlation coefficient measurement and used to measure the correlation of the repeated measurements here. Reliability is quantified in anthropometric studies using the Technical Error of Measurement (TEM) [15], which is essentially a form of standard deviation, and reliability coefficient. The TEM, or called r, was calculated using the following equation

$$r = \sqrt{\frac{\sum_{i=1}^{n} (\sum_{j=1}^{k} x_{j}^{2} - (\frac{1}{k})(\sum_{j=1}^{k} x_{j})^{2})}{n(k-1)}}$$
(17)

where x_j^2 is the squared value of the j_{th} replicate (j=1,2,...,k), n is the number of subjects, and k is the number of measurements per subject. The reliability coefficient (R) [15] is calculated due to the measurement error (r^2) against the sample variance (s^2) and calculated with follow expression

$$R = 1 - (r^2/s^2). (18)$$

The TEM and *R* of 2D image-based measurement in precision experiment are shown in TABLE V.

According to the definition of reliability coefficient, if the technical measurement error is much smaller than the sample variance, the reliability coefficient of measurement will be high. As the show in TABLE V, the reliability coefficients of 2D image-based measurement are all above 99 % and proofed that the 2D image-based system has highly reliability and repeatable.

TABLE V. THE TEM AND RELIABILITY COEFFICIENTS OF 2-D IMAGE-BASED MEASUREMENT

Measurement	TEM (r)	R
Shoulder length	0.11	99.73%
Chest circumference	0.32	99.41%
Waist circumference	0.37	99.50%
Hip circumference	0.29	99.29%
Leg length	0.28	99.39%

VI. CONCLUSION

This paper introduces a new approach to anthropometric measurement based on 2D image, and the results of experiment show that the approach is quite comparable to traditional measurement method performing by skilled anthropometrists, both in terms of accuracy, precision and repeatability. The advantage of using this approach is that the people can easy to take an anthropometric measurement by using digital camera, which is very popular today, without heavy and complicated setting, and the high performance of accuracy and reliability compared to manual measurement could be obtained. Furthermore, the proposed method can provide the measurement for people regardless of where, when, or by whom.

For many applications, such as large scale anthropometric investigation of population, it can be done by using the proposed system, which has high performance of accuracy and low cost. In the future, the new type of apparel merchandising would be proposed that the customer can buy the clothes at home through internet and less return for unfitting reason by using 2D measurement in advance. Furthermore, the customized clothe could be made by only using photographs and without he or she physical presence.

REFERENCES

- C. C. Gordon, T. Churchill, C. E. Clauser, B. Bradtmiller, J. T. McConville, I. Tebbetts, and R. A. Walker, 1988 anthropometric survey of US army personnel: methods and summary statistic. NATICK/TR-89/044. US Army Natick Research, Development, and Engineering Center, Natick, MA, 1989.
- [2] P. Meunier, and S. Yin, "Performance of a 2D image-based anthropometric measurement and clothing sizing system," Applied Ergonomics, vol. 31, pp. 445-451, October 2000.
- [3] C. Y. Hung, P. Witana, and S. Goonetilleke, "Anthropometric measurements from photographic images," 7th Int. Proc. on Work with computing system, pp. 104-109, June 2004.
- [4] E. Paquet and H. L. Viktor, "Adjustment of Virtual Mannequins Through Anthropometric Measurements, Cluster Analysis, and Content-Based Retrieval of 3-D Body Scans," IEEE Trans. on Instrumentation and Measurement, vol. 56, pp.1924 - 1929, October 2007.
- [5] K. Robinette, H. Daanen, and E. Paquet, "The CAESAR project: A 3-D surface anthropometric survey," 2nd Int. Conf. on 3-D digital imaging and modeling, Ottawa, Ont., Canada, pp. 380-386, October 1999.
- [6] D. Burnsides, M. Boehmerk, and K. Robinette, "3-D landmark detection and identification in the CAESAR project," 3rd Int. Conf. on 3-D Digital imaging and modeling, Quebec City, Que., Canada, pp. 393-398, May 2001.
- [7] J. F. David and K. Y. Yan, "Automatic image segmentation by integrating color-edge extraction and seeded region growing," IEEE Trans. on image processing, vol. 10, pp. 1454-1466, October 2001.
- [8] A. Albiol, L. Torres, and E. J. Delp, "Optimum color spaces for skin detection," IEEE Int. Conf. Image processing, Thessaloniki, Greece, vol. 1, pp. 122-124, October 2001.
- [9] P. J. Shen and K. W. Wu, Massage For Pain Relief: A Step-by-Step Guide, Morning star, Taipei, 2003.
- [10] A. Watt and H. Watt, Advanced Animation and Rendering Techniques: Theory and Practice, Addison Wesley, NY, 1992.
- [11] Taiwan Human Body Bank (TAIBBK), http://3d.cgu.edu.tw/DesktopDefault.asp, accessed September. 2009.

- [12] C. Y. Yu, Y. H. Lo and W. K. Chiou, "The 3D Scanner for Measuring BodySurface Area: A Simplified Calculation in the Chinese Adult," Applied Ergonomics, Vol.34, pp.273-278, 2003.
- [13] C. T. Lin and C. S. Lee, Neural Fuzzy Systems: A Neuro-Fuzzy Synergism to Intelligent System, Prentice Hall International, 1996.
- [14] A. Chamberland, R. Carrier, F. Forest, and G. Hachez, Defence and Civil Institute of Environmental Medicine, Toronto, Ontario, 1997.
- [15] T. G. Loman, A. F. Roche, and R. Martorell, Anthropometric standardization reference manual, Human Kinetics Books, Champaign, IL, 1988.

AUTHORS PROFILE

Sheng-Fuu Lin was born in Tainan, R.O.C., in 1954. He received the B.S. and M.S. degrees in mathematics from National Taiwan Normal University in 1976 and 1979, respectively, the M.S. degree in computer science from the University of Maryland, College Park, in 1985, and the Ph.D. degree in electrical engineering from the University of Illinois, Champaign, in 1988. Since 1988, he has been on the faculty of the Department of Electrical and Control Engineering at National Chiao Tung University, Hsinchu, Taiwan, where he is currently a Professor. His research interests include image processing, image recognition, fuzzy theory, automatic target recognition, and scheduling.

Shih-Che Chien was born in Chiayi, R.O.C., in 1978. He received the B.E. degree in electronic engineering from the Nation Chung Cheng University, in 2002. He is currently pursuing the M.E. and Ph.D. degree in the Department of Electrical and Control Engineering, the National Chiao Tung University, Hsinchu, Taiwan. His current research interests include image processing, image recognition, fuzzy theory, 3D image processing, intelligent transportation system, and animation.

Kuo-Yu Chiu was born in Hsinchu, R.O.C., in 1981. He received the B.E. degree in electrical and control engineering from National Chiao Tung University, Hsinchu, Taiwan, R.O.C., in 2003. He is currently pursuing the Ph. D. degree in the Department of Electrical and Control Engineering, the National Chiao Tung University, Hsinchu, Taiwan. His current research interests include image processing, face recognition, face replacement, intelligent transportation system, and machine learning.

A fast fractal image encoding based on Haar wavelet transform

Sofia Douda Département de Mathématiques et Informatique & ENIC, Faculté des Sciences et Techniques, Université Hassan 1^{er}, Settat, Morocco.

Abdallah Bagri ENIC, Faculté des Sciences et Techniques, Université Hassan 1^{er}, Settat, Morocco. Abdelhakim El Imrani LCS, Faculté des Sciences, Université Mohammed V, Rabat, Morocco.

Abstract—In order to improve the fractal image encoding, we propose a fast method based on the Haar wavelet transform. This proposed method speed up the fractal image encoding by reducing the size of the domain pool. This reduction uses the Haar wavelet coefficients. The experimental results on the test images show that the proposed method reaches a high speedup factor without decreasing the image quality.

Keywords- Fractal image compression, PIFS, Haar wavelet transform, SSIM index.

I. INTRODUCTION

Fractal image compression (FIC) is one of the recent methods of image compression firstly presented by Barnsley and Jacquin [1-5]. This method is characterized by its high compression ratio which is achieved with an acceptable image quality [6], a fast decoding and a multi-resolution property. It is based on the theory of Iterated Function System (IFS) and on the collage theorem. Jacquin [3-5] developed the first algorithm of FIC by Local or Partitioned Iterated Function Systems (PIFS) which makes use of local self-similarity propriety in real images. In FIC, the image is represented through a contractive transformation defined by PIFS for which the decoded image is approximately its fixed point and close to an input image.

In Jacquin's algorithm, an input image is partitioned into non-overlapping sub-blocks R_i called range blocks, the union of which covers the whole image. Each range block R_i is put in corresponding transformation with another part of a different scale, called domain block, looked for in the image. The domain blocks can be obtained by sliding a window of the same size around the input image to construct the domain pool. The classical encoding method, i.e. full search, is time consuming because for every range block the corresponding block is looked for among all the domain blocks. Several methods are proposed to reduce the time encoding. The most common approach is the classification scheme [6-10]. In this scheme, the domain and the range blocks are grouped in a number of classes according to their common characteristics. For each range block, comparison is made only for the domain blocks falling into its class. Fisher's classification method [6] constructed 72 classes for image blocks according to the variance and intensity. In Wang et al. [10], four types of range

blocks were defined based on the edge of the image. Jacobs et al. uses skipping adjacent domain blocks [11] and Monro and Dudbridge localizes the domain pool relative to a given range block based on the assumption that domain blocks close to this range block are well suited to match the given range block [12]. Methods based on reduction of the domain pool are also developed. Saupe's Lean Domain Pool method discards a fraction of domain blocks with the smallest variance [13] and in Hassaballah et al., the domain blocks with high entropies are removed from the domain pool [14]. Other approaches focused on improvements of the FIC by tree structure search methods [15, 16], parallel search methods [17, 18] or using two domain pools in two steps of FIC [19]. The spatial correlation in both the domain pool and the range pool was added to improve the FIC as developed by Truong et al. [20]. Tong [21] proposes an adaptive search algorithm based on the standard deviation (STD). Other approaches based on genetic algorithms are also applied to speed up the FIC [22-23]. In these methods, higher speedup factor are often associated with some loss of reconstructed image quality. In the present work, a new method is proposed to reduce the encoding time of FIC using the Haar wavelet transform. It speeds up the time encoding by discarding the smooth domain blocks from the domain pool. The type of these blocks is defined using the Haar wavelet transform. A high speedup factor is reached and the image quality is still preserved.

II. THE PROPOSED METHOD BASED ON HAAR WAVELET TRANSFORM

A. The Haar wavelet transform

The Haar Wavelet Transform (HWT) [24] is one of the simplest and basic transformations from the space domain to a local frequency domain and it is a very useful tool for signal analysis and image processing. The HWT decompose a signal into different components in the frequency domain. One-dimensional HWT decomposes an input sequence into two components (the average component and the detail component) by applying a low-pass filter and a high-pass filter. In the HWT of 2D image of size NxN, a pair of low-pass and high-pass filters is applied separately along the horizontal and vertical direction to divide the image into four sub-bands of size N/2xN/2 (Fig. 1). After one level of decomposition, the low-

low-pass sub-band LL is the multiresolution approximation of the original image, and the other three are high frequency subbands representing horizontal, vertical and diagonals edges, respectively. The LL band is again subject to the same procedure.



Figure 1. The result of 2D image HWT decomposition.

This wavelet decomposition can be repeatedly applied on the low-low-pass sub-band at a coarser scale unless it only has one component as shown in fig. 2.

Let D be a given image block of size NxN. D can be decomposed into one component low-pass signal by a log N/log 2 pyramidal HWT. In the case of N=8, D can be transformed to one component by 3 decomposition (Fig. 2). The LL³ band in level 3 is the multiresolution approximation of LL² bands in level 2. The coefficients HL³, LH³ and HH³ of the highest level denote the coarsest edges along horizontal, vertical and diagonal directions respectively in level 2.

LL ³	HL^3 HH^3	HL^2	HL^1
L	H^2	HH^2	TIL
	LH ¹		HH^1

Figure 2. The result of three level HWT pyramidal decomposition of an image block of size 8x8.

We will refer to these coefficients obtained at the highest level hereafter as WH_D for HL^3 , WV_D for LH^3 and HH^3 for WD_D .

If both WH_D and WV_D are small, then the block D tends to have less edge structure (smooth block). When a block has high degree of edge structure, either WH_D or WV_D will be large. If WH_D is larger, D will have horizontal edge properties. On the other hand, if WV_D is larger, then D will have vertical edge properties. Finally, those blocks with high magnitudes of WH_D and/or WV_D are designed as heterogeneous domain blocks. The type of each block D is determined as follows:

$$\begin{aligned} &\text{if } \left| w_{H_D} \right| < T_w \text{ and } \left| w_{V_D} \right| < T_w \\ &\text{then D is a smooth domain block} \end{aligned} \tag{1}$$
 else D is a heterogeneous domain block

where I.I denotes the absolute value of its variable and $T_{\rm W}$ is a threshold.

Thus, an image block D can be determined as belonging to smooth or heterogeneous type by using its vertical coefficient WV_D and its horizontal coefficient WH_D obtained by a pyramidal HWT at the highest level.

The computation of WH_D and WV_D do not require the calculation of other wavelets coefficients. Indeed, let D be an image block of size 4x4 represented as follows:

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16

Then, analyzing the expression of the coefficients WH_D and WV_D obtained after two pyramidal HWT, allow us to find the following simplified formula:

$$WH_D = \frac{A - B}{16} \tag{2}$$

$$WV_D = \frac{C - D}{16} \tag{3}$$

where A = 1 + 2 + 5 + 6 + 3 + 4 + 7 + 8, B = 9 + 10 + 13 + 14 + 11 + 12 + 15 + 16,

C=1+2+5+6+10+13+14, D=3+4+7+8+11+12+15+16.

B. The proposed method

The proposed method is aimed to reduce the encoding time by reducing the cardinal of the domain pool. As only a fraction of the domain pool is used in fractal encoding and the set of the used blocks is localized along edges and in the regions of high contrast of the image (designed as heterogeneous blocks), it's possible to reduce the cardinal of the domain pool by discarding the smooth domain blocks. Therefore, each range bloc is compared only to the heterogeneous domain blocks. This method of reduction of the domain pool is simple since only few computations are required to calculate the coefficients WHD and WVD of a domain block D to classify it as smooth or heterogeneous.

The threshold T_W can be fixed or chosen in an adaptive way. Determining T_W adaptively allow us to choose the speedup ratio. The main idea is to set the thresholds such that a fraction α of the domain pool can be discarded. The value of α can be one third, tow thirds,... of the domain pool. Due to the fact that the encoding time depends on the number of comparisons between range and domain blocks, the speedup ratio can be estimated.

The determination of the threshold T_W , which depends on the fraction α of the domain pool to be eliminated, is summarised as follows:

- For each domain block D, calculate the Haar wavelet coefficient WH_D and WV_D . Set $S_D = max(|WH_D|, |WV_D|)$.
- Sort all the values of S_D in increasing order.

Find S^* corresponding to the value of α . Set the threshold $T_w = S^*$.

Due to the fact that we apply our method in the case of a quadtree partitioning, we choose different thresholds for every size of the domain blocks.

The first steps of the proposed method are as follows:

- Choose a value of α .
- Construct the domain pool.
- Compute the Haar wavelet coefficients WHDI and |WV_D| for each domain block D.
- Determine the threshold T_W for each size domain
- Remove from the domain pool the smooth domain blocks.

III. **EXPERIMENTAL RESULTS**

The different tests are performed on three 256x256 images, represented in fig. 3, with 8 bpp on PC with Intel Pentium Dual 2.16 Ghz processor and 2 GO of RAM. The quadtree partitioning [6] is adopted for the FIC. The encoding time is measured in seconds. The rate of compression is represented by the compression ratio (CR), i.e. the size of the original image divided by the size of the compressed image. The speedup factor (SF) of a particular method can be defined as the ratio of the time taken in full search to that of the said method, i.e.,

$$SF = \frac{\text{Time taken in full search}}{\text{Time taken in a particular method}}$$
 (4)

The image quality is measured by the peak signal to noise ratio (PSNR) and the structural similarity Measure (SSIM) index [25].

The PSNR of two images X and Y of sizes N, is defined as follows:

$$PSNR = 10 \times \log \left(\frac{255^2}{MSE} \right)$$
 (5)

where

MSE =
$$\frac{1}{N} \sum_{i=1}^{N} (x_i - y_i)^2$$
 (6)

x_i and y_i are the gray levels of pixel of the original image and the distorted image respectively.

The SSIM index is a method for measuring the similarity between two images x and y defined by Wang [25] as follows:

$$SSIM(x,y) = \frac{(2\mu_x \mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$
(7)

where
$$\mu_x = \frac{1}{N} \sum x_i$$
, $\mu_y = \frac{1}{N} \sum y_i$, $\sigma_x = (\frac{1}{N-1} \sum (x_i - \mu_x)^2)^{\frac{1}{2}}$,

$$\sigma_y = (\frac{1}{N-1} \sum (y_i - \mu_y)^2)^{\frac{1}{2}} , \ \sigma_{xy} = \frac{1}{N-1} \sum (x_i - \mu_x)(y_i - \mu_y)) \ .$$

C1 and C2 are positive constants chosen to prevent unstable measurement when $(\mu_x^2 + \mu_y^2)$ or $(\sigma_x^2 + \sigma_y^2)$ is close to zero. They are defined in [25] as:

$$C_1 = (K_1 L)^2$$
, $C2 = (K_2 L)^2$ (8)

where L is the dynamic range of pixel values (L= 255 for 8-bit gray scale images). K_1 and K_2 are the same as in [20]: K1=0.01 and K2 = 0.03.

In the present work, we use a mean SSIM (MSSIM) index to evaluate the overall image quality:

MSSIM(X,Y) =
$$\frac{1}{M} \sum_{i=1}^{M} SSIM(x_i, y_i)$$
 (9)

where X and Y are the original and the distorted images respectively; x_i and y_i are the image contents at the ith local window of size 8x8 and M is the number of local windows of the image.







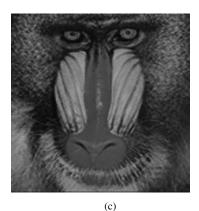


Figure 3. Images of size 256 x 256 : Lena (a), Peppers (b) and baboun (c).

Table 1 and 2 gives the encoding time, the compression ratio, the image qualities and the speedup factor measured on the three test images for different values of $(1-\alpha)$. The full search occurs when $\alpha=1$ and there is no time reduction because no domain block is eliminated. The results show that the encoding time scales linearly with α as illustrated in fig. 4. For values of α between 0.9 and 0.3, there is no degradation in the image quality. On contrary, the PSNR improves slightly for the test images. The SF of 10 causes a drop of PSNR of 0.6 dB, 0.55 dB and 0.07 dB for Lena, Peppers and Baboun images respectively.

TABLE I. THE RESULTS OF THE PROPOSED METHOD FOR LENA AND PEPPERS IMAGES.

			Lena			Peppers				
α	Time	CR	PSNR	MSSIM	SF	Time	CR	PSNR	MSSIM	SF
1	19.64	10.46	30.92	0.8909	1.00	19.64	10.98	31.91	0.8931	1.00
0.9	18.17	10.41	30.94	0.8915	1.08	16.71	10.98	31.91	0.8931	1.18
0.8	15.32	10.35	30.98	0.8933	1.28	15.18	10.94	31.93	0.8936	1.29
0.7	14.13	10.30	30.99	0.8934	1.39	13.85	10.92	31.92	0.8936	1.42
0.6	11.72	10.19	31.05	0.8948	1.68	11.40	10.81	31.94	0.8938	1.72
0.5	10.30	10.11	31.05	0.8947	1.91	9.86	10.65	31.98	0.8958	1.99
0.4	7.99	9.90	31.06	0.8971	2.46	7.71	10.49	31.99	0.8956	2.55
0.3	6.40	9.66	31.02	0.8957	3.07	6.05	10.45	31.88	0.8943	3.25
0.2	4.31	9.47	30.88	0.8936	4.56	4.09	10.20	31.77	0.8927	4.80
0.1	2.39	9.09	30.48	0.8891	8.22	2.29	9.88	31.47	0.8877	8.58
0.08	1.92	9.03	30.32	0.8872	10.23	1.94	9.68	31.36	0.8870	10.12
0.06	1.56	9.02	30.13	0.8843	12.59	1.51	9.54	31.13	0.8839	13.01
0.04	1.11	8.81	29.89	0.8805	17.69	1.14	9.29	30.90	0.8811	17.23
0.02	0.67	8.73	29.53	0.8737	29.31	0.70	9.05	30.43	0.8752	28.06
0.008	0.45	8.63	29.24	0.8683	43.64	0.47	8.82	29.95	0.8690	41.79
0.006	0.39	8.50	29.13	0.8655	50.36	0.41	8.76	29.56	0.8631	47.90
0.004	0.36	8.33	29.03	0.8637	54.56	0.36	8.80	29.06	0.8567	54.56

TABLE II. THE RESULTS OF THE PROPOSED METHOD FOR BABOUN IMAGE.

α			Babou	n	
"	Time	CR	PSNR	MSSIM	SF
1	25.47	7.46	32.55	0.8802	1.00
0.9	22.29	7.45	32.55	0.8805	1.14
0.8	21.54	7.45	32.56	0.8805	1.18
0.7	17.07	7.40	32.57	0.8813	1.49
0.6	14.71	7.33	32.63	0.8829	1.73
0.5	12.95	7.33	32.64	0.8831	1.97
0.4	10.84	7.27	32.65	0.8828	2.35
0.3	7.74	7.22	32.66	0.8827	3.29
0.2	5.40	7.11	32.65	0.8829	4.72
0.1	2.87	6.90	32.57	0.8806	8.87
0.08	2.31	6.81	32.48	0.8785	11.03
0.06	1.72	6.73	32.43	0.8769	14.81
0.04	1.39	6.66	32.24	0.8741	18.32
0.02	0.84	6.42	31.95	0.8693	30.32
0.008	0.56	6.29	31.67	0.8632	45.48
0.006	0.47	6.21	31.22	0.8533	54.19
0.004	0.45	6.22	31.18	0.8523	56.60

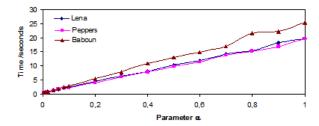


Figure 4. Effect of parameter α on encoding time.

The CR decreases slightly when SF \leq 0.2 (0.99 for Lena, 0.78 for Peppers and 0.35 for Baboun). When SF increases, the CR decreases. A higher SF is accompanied with a high decrease of CR. This could be explained by the fact that some large range blocks could be covered well by some domain blocks which were excluded from the domain pool. Therefore, these large range blocks are subdivided in four quadrants resulting in a decrease of CR. For example, when SF \approx 42 the drops of CR are 1.83, 2.16 and 1.17 for Lena, Peppers and Baboun respectively. For comparison, the full search reaches a PSNR of 30.92 dB with a required time of 19.64 seconds for Lena image. In the proposed method, the encoding time of Lena image is 1.11 seconds while the PSNR is 29.89 dB when α =0.04. The speedup factor attains 17.69 with a drop of PSNR of 1.03.

For visual comparison, fig. 5, fig. 6 and fig. 7 shows examples of reconstructed images encoded using full search and the proposed method.



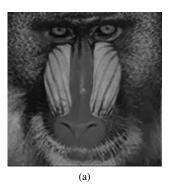


Figure 5. Reconstructed image Lena by full search (a) and by the proposed method (b) when SF = 17.69.





Figure 6. Reconstructed image Peppers by full search (a) and by the proposed method (b) when SF = 17.23.



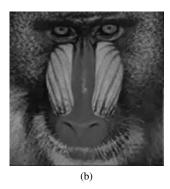
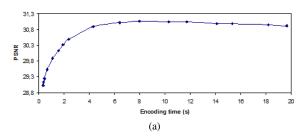


Figure 7. Reconstructed image Baboun by full search (a) and by the proposed method (b) when SF = 18.32.

Fig. 8, fig. 9 and fig. 10 show that PSNR and MSSIM vary in the same way according to the encoding time for the test images.



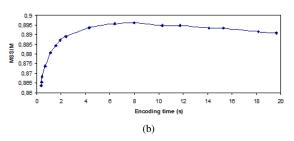


Figure 8. Encoding time versus PSNR (a) and MSSIM (b) for Lena.

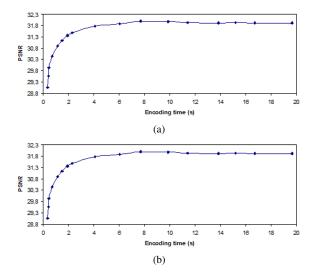


Figure 9. Encoding time versus PSNR (a) and MSSIM (b) for Peppers.

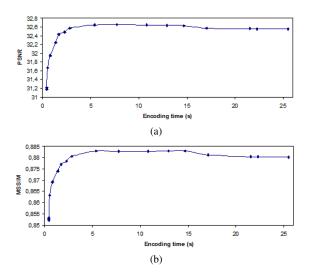


Figure 10. Encoding time versus PSNR (a) and MSSIM (b) for Baboun.

When comparing our method to Hassaballah et al. method (HM), we find that our method preserve well the image quality and mark a slight decrease of CR for a high speedup factor than HM. For Lena image, a speedup factor of 10.23 is reached with a drop of 0.6 dB and a decrease of the CR by 1.43 while HM cause a drop of PSNR of 3.69 dB and a drop of CR of 3.28. For the same SF, the drop of MSSIM is of 0.024 by our method and 0.0668 by HM. Furthermore, the results of encoding are still better than HM when the SF achieve a high value for the three test images.

The comparison with Saupe's method (SM) shows an improvement of the encoding time without drops of PSNR, MSSIM and CR. For example, with Lena image, a SF of 29.31 results in a PSNR of 29.53 dB, a CR of 8.73 and a MSSIM of 0.8737. While by SM a SF of 21.81 generate a PSNR of 29.30 dB, a CR of 8.51 and a MSSIM of 0.8654.

Also the comparison with AP2D-ENT [26] shows that the proposed method reaches higher SFs with lesser drops of PSNR and of MSSIM than AP2D-ENT. Indeed for Lena image, a SF of 43.64 generates a PSNR of 29.24 dB, a CR of 8.63 and a MSSIM of 0.8683. While the SF of 24.76 obtained by AP2D-ENT, generates a PSNR of 28.63 dB, a CR of 8.84 and MSSIM of 0.8507. Similar improvements are observed for Peppers and Baboun images.

IV. CONCLUSION

In this study, we propose to reduce the time of fractal image encoding by using a new method based on the Haar wavelet transform (HWT). The two HWT horizontal and vertical coefficients obtained at the last level of pyramidal decomposition are used to determine the smooth or heterogeneous type of domain blocks. The proposed method reduces the encoding time by removing the smooth domain blocks from the domain pool. Experimental results show that discarding a fraction of smooth blocks has little effect on the image quality while a high speedup factor is reached.

REFERENCES

- M. F. Barnsley and A. D. Sloan, "A better way to compress images", BYTE magazine, pp. 215-223, 1988.
- [2] M. F. Barnsley, "Fractal every where". New-york: Academic Press, California, 1988.
- [3] A. E. Jacquin, "A fractal theory of iterated Markov operators on spaces of measures with applications to digital image coding", PhD Thesis, Georgia Institute of Technology, 1989.
- [4] A. E. Jacquin, "A novel fractal block coding technique for digital image", IEEE Int.Conf. on ASSP, ICASSP-90, pp. 2225-2228, 1990.
- [5] A. E. Jacquin, "Image coding based on a fractal theory of iterated contractive image transformations", IEEE Trans. on Image Processing, Vol. 1, pp.18-30, January 1992.
- [6] Y. Fisher, "Fractal Image Compression: Theory and Application", Springer-verlag, New York, 1994.
- [7] D. J. Duh, J. H. Jeng and S. Y. Chen, "DCT based simple classification scheme for fractal image compression", Image and vision computing, Vol. 23, pp. 1115-1121, 2005.
- [8] X. Wu, D. J. Jackson and H. Chen, "A fast fractal image encoding method based on intelligent search of standad deviation", Computers and Electrical Engineering, Vol. 31, pp. 402-421, 2005.
- [9] T. Kovacs, "A fast classification based method for fractal image encoding", Image and Vision Computing, Vol. 26, pp. 1129-1136, 2008.
- [10] Z. Wang, D. Zhang and Y. Yu, "Hybrid image coding based on partial fractal mapping", Signal Process: Image Commun., Vol. 15. pp. 767-779, 2000.
- [11] E. W. Jacobs, Y. Fisher and R. D. Boss, "Image compression: A study of iterated transform method", Signal process, Vol. 29, pp. 251-263, 1992.
- [12] D. M. Monro and F. Dudbridge, "Approximation of image blocks", In Proc. Int. Conf. Acoustics, Speed. Signal Processing, Vol. 3, pp. 4585-4588, 1992.
- [13] D. Saupe, "Lean domain pools for fractal image compression", Journal of Electronic Imaging, Vol. 8, pp. 98-103, 1999.
- [14] M. Hassaballah, M. M. Makky and Y. B. Mahdi. A fast fractal image compression method based entropy. Electronic Letters on Computer Vision and Image Analysis, Vol. 5, pp. 30-40, 2005.
- [15] B. Bani-Eqbal, "Enhancing the speed of fractal image compression, Optical Engineering", Vol. 34, No. 6, pp. 1705-1710, 1995.

- [16] B. Hurtgen and C. Stiller, "Fast hierarchical codebook search for fractal coding still images", in Proc. EOS/SPIE Visual Communications PACS Medical Applications, Vol. 1977, pp. 397-408, 1993.
- [17] C. Hufnagel and A. Uhl, "Algorithms for fractal image compression on massively parallel SIMD arrays", Real-Time Imaging, Vol. 6, pp. 267-281, 2000.
- [18] D. Vidya, R. Parthasarathy, T. C. Bina and N. G. Swaroopa, "Architecture for fractal image compression". J. Syst. Archit., Vol. 46, pp. 1275-1291, 2000.
- [19] S. Douda, A. El Imrani and M. Limouri, "Une nouvelle approche d'accélération du codage fractal d'images", ARIMA, Vol. 11, pp. 97-114, 2009.
- [20] T. K. Troung, C. M. Kung, J. H. Jeng and M. L. Hsieh, "Fast fractal image compression using spatial correlation", Chaos Solitons & Fractals, Vol. 22, pp. 1071-1076, 2004.
- [21] C. S. Tong, M. Pi, "Fast fractal image encoding based on adaptive search", IEEE Trans Image Process, Vol. 10, pp.1269-1277, 2001.

- [22] M.-S. Wu, J.-H. Jeng and J.-G. Hsieh, "Schema genetic algorithm for fractal image compression", Eng. Appl. Artif. Intell. Vol. 20, No 4, pp. 531-538, 2007.
- [23] M.-S. Wu and Y.-L. Lin, "Genetic algorithm with hybrid select mechanism for fractal image compression", Digital Signal Process, Vol. 20, No 4, pp. 1150-1161, 2010.
- [24] S.G.Mallat, "A theory of multiresolution signal decomposition: the wavelet representation", IEEE Trans. PAMI, vol. 11, pp. 674-693, July 1989.
- [25] Z. Wang, A. Bovik, H. Sheikh and E. Simoncelli. "Image quality assessment: From error visibility to structural similarity". IEEE Transactions on Image Processing, Vol. 13, No. 4, pp. 600–612, 2004.
- [26] S. Douda, A. El Imrani and A. Bagri, "A new approach for improvement of fractal image encoding", IJCSE, Vol. 02, N°. 04, pp. 1387-1394, 2010

A New Noise Estimation Technique of Speech Signal by Degree of Noise Refinement

Md. Ekramul Hamid College of Computer Science King Khalid University Abha, Kingdom of Saudi Arabia Md. Zasim Uddin Dept. of Computer Science University of Rajshahi Rajshahi, Bangladesh. Md. Humayun Kabir Biswas College of Computer Science King Khalid University Abha, Kingdom of Saudi Arabia Somlal Das Dept. of Computer Science University of Rajshahi Rajshahi, Bangladesh.

Abstract— An improved method for noise estimation of speech utterances which are disturbed by additive noise is presented in this paper. Here, we introduce degree of noise refinement of minima value sequence (MVS) and some additional techniques for noise estimation. Initially, noise is estimated from the valleys of the spectrum based on the harmonic properties of noisy speech, called MVS. However, the valleys of the spectrum are not pronounced enough to warrant reliable noise estimates. We, therefore, initially use the estimated Degree of Noise (DON) to adjust the MVS level. For every English phoneme DON is calculated and averaged within those processing frames for the each input SNR. We consider this calculated average DONs as standard value corresponding to the input SNR which is aligned with the true DON using the least-squares (LS) method results a function to estimate the degree of noise. Therefore, using the technique, it is possible to estimate the state of the added noise more accurately. We use two stage refinements of estimated DON to update the MVS as well as to estimate a nonlinear weight for noise subtraction. The performance of the proposed noise estimation is good when it is integrated with the speech enhancement technique.

Keywords-component; Noise Estimation, the Degree of Noise, Speech Enhancement, Nonlinear Weighted Noise Subtraction

I. INTRODUCTION

Noise estimation is one of the most important aspects for single channel speech enhancement. Usually in single-channel speech enhancement systems, most algorithms require a voice activity detector (VAD) and the speech/pause detection plays the major role in the performance of the whole system. However, these systems can perform well for voiced speech and high signal-to-noise ratio (SNR), but their performance degrades with unvoiced speech in low SNR.

Traditional noise estimators are based on voice activity detectors (VAD) which are difficult to tune and their application to low SNR speech results often in clipped speech. The original MMSE-STSA estimates the noise power spectrum on the basis of the noisy speech only in the first non-speech period where the pure noise is available [1]. However, these systems can perform well only for voiced speech and high SNR. Martin (2001) proposed a method for estimating the noise spectrum based on tracking the minimum of the noisy. The main drawback of this method is it fails to update the noise spectrum when the noise floor increases abruptly [2]. Cohen (2002) [3] proposed a minima controlled recursive algorithm (MCRA) which updates the noise estimate by tracking the

noise-only regions of the noisy speech spectrum. In the improved MCRA approach (Cohen 2003) [4], a different method was used to track the noise-only regions of the spectrum based on the estimated speech-presence probability. Doblinger [5] updated the noise estimate by continuously tracking the minimum of the noisy speech in each frequency bin. As such, it is computationally more efficient than the method in Martin 2001. However, it fails to differentiate between an increase in noise floor and increase in speech power. Hirsch and Ehrlicher [6] updated the noise estimate by comparing the noisy speech power spectrum to the past noise estimate. This method fails to update the noise when the noise floor increases abruptly and stays at that level. In our previous study, Hamid (2007) [7] proposed the noise estimation by using the MVS. The noise floor is updated with the help of estimated DON. Here DON is estimated on the basis of pitch and the pitch of unvoiced sections is not accurately estimated.

In this paper, we propose a method which has good noise tracking and controlling capability. To estimate noise, first we search for the valleys of the amplitude spectrum on a frame by frame basis and estimate minima values of the spectrum, called minima value sequence (MVS). To improve the estimation accuracy of MVS, we use DON. As it is a single-channel method, direct estimation of the degree of noise is not possible. For that, frame wise averaged DON is estimated from the estimated noise of the observed signal. We have considered these DONs as standard value corresponding to the input SNR. Then each of these estimated 1st averaged DONs for corresponding input SNR is aligned with the true DON using the least-squares (LS) method results the 1st estimated degree of noise (DON1) of that frame. The 1st estimated DON1 is applied to update the MVS. Next, the noise level is reestimated and from the estimated noise, we again estimate 2nd averaged DON and similarly get the 2nd estimated DON2. We used the 2nd estimated DON2 to estimate the weight for noise subtraction process. Because noise is estimated from the estimated DONs, which is obtained from the true DON, so it is possible to estimate noise amplitudes in more accurate form with lower speech distortion and able to suppress musical noise in the enhanced speech.

II. PROPOSED NOISE ESTIMATION METHOD

We have assumed that speech and noise are uncorrelated to each other. Let y(n)=s(n)+d(n), where y(n) is the observed noisy speech signal, s(n) is the clean speech signal and d(n)

is the additive noise. We further assume that signal and noise are statistically independent. Under the above assumptions, we can write the powers as $P_v = P_s + P_d$.

A. Estimation of the minima value sequence (MVS)

The sections of consecutive samples are used as a single frame l (320 samples). Consecutive frames are spaced l' (100 samples) achieving an almost 62.75% overlap between them. The short-term representation of a signal y(n) is obtained by windowing (Hamming window) and analyzed using N=512 point discrete-Fourier transform (DFT) in sampling frequency 16KHz. Initially, noise spectrum is estimated from the valleys of the amplitude spectrum and we assume that the peaks correspond to voice parts and valleys are the noise only parts. The algorithm for noise estimation is as follows:

- 1. Compute the RMS value Y_{rms} of the amplitude spectrum Y(k). We detect the minima $Y_{\min}(k_{\min}) \leftarrow \min(Y(k))$ values of Y(k) when the following condition (Y(k) < Y(k-1)) and Y(k) < Y(k+1) and $Y(k) < Y_{rms}$ is satisfied. The k_{\min} expresses the positions of the frequency bin index of minima values.
- 2. Interpolate between adjoining minima positions $(k_{\min} \leftarrow k)$ to obtain the minimum value sequences (MVS) $Y_{\min}(k)$.
- We smooth the sequences by taking partial average called smoothed minimum value sequences (SMVS).
 This process continuously updates the estimation of noise among every analysis frames.

An estimation of noise from the SMVS is survived by an overestimation and underestimation of the SNR. To achieve good tracking capability with controlled overestimation problem, the proposed noise estimation algorithm adopting the concept of DON. The block diagram of the noise estimation process is given in Figure 1.

B. Estimation of the Degree of Noise (DON)

In a single-channel method, we only know the power of the observed signal. Therefore, direct estimation of the degree of noise (P_d/P_{obs}) is not possible. For that, frame wise DON is estimated from the estimated noise of the observed signal of each frame m. For optimal estimation of DON, we are carried out our experiment on 20 vowel phonemes of 3 male and 3 female taken from TIMIT database. First white noise of various SNR are added to these voiced vowel phonemes. Then for each SNR white noisy phonemes are processed frame wise and DON is estimated in each frame for each phoneme individually. For every phoneme DON is averaged within those processing frames for the corresponding input SNR. Then each of these estimated 1^{st} averaged DONs of each frame m for corresponding input SNR expressed as \bar{Z}_{lm} . This \bar{Z}_{lm} is aligned with the true DON (Ztr) using the least-squares (LS) method results the 1st estimated degree of noise (DON1) Z_{lm} of that frame. The true DON (Z_{tr}) is given by

$$Z_{tr} = \frac{P_d}{P_s + P_d} = \frac{1}{1 + 10^{\frac{dB}{10}}} \tag{1}$$

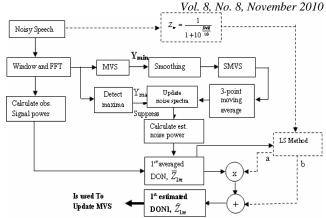


Figure 1. Block diagram of the 2nd estimated DON1, Z_{lm}

 1^{st} averaged estimated DON $\overline{Z}_{\text{\tiny lm}}$

$$\overline{Z}_{1m} = \frac{1}{M} \sum_{m=1}^{M} \frac{P_{\eta}(m)}{P_{obs}(m)}$$
(2)

where, M are the noise added frames; $P_{\eta}(m)$ and $P_{obs}(m)$ are the powers of noise and observed signals, respectively. Here it obvious that we consider only the voiced phonemes in our experiment. So the averaged DON value should be limited to voiced portion of a speech sentence. But practically the unvoiced portion contaminated with higher degree of noise. Hence the estimated noise is higher for unvoiced frame than from voiced frame. Consequently higher DON value is obtained from unvoiced frame than from voiced frame that is logically resemblance. The degree of noise estimated from a previously prepared function using least square method is given by [7]

$$Z_{1m} = a \times \overline{Z}_{1m} + b \tag{3}$$

where Z_{lm} is the 1st estimated DON1 of frame m. The error between the true and the estimated values can be minimized by tuning a, b. In the experiment, 20 phoneme sounds for 3 male and 3 female degraded by the white noise in different SNRs (-10,-5,0,...,30 dB) is considered. The value of Z_{lm} is applied to update the MVS. Next, the noise level is re-estimated with the help of Z_{lm} . Finally, from the estimated noise, we again estimate 2^{nd} averaged DON (\overline{Z}_{2m}) and similarly the 2^{nd} estimated DON2 (Z_{2m}) which is used to estimate the noise weight for nonlinear weighted noise subtraction.

We conduct an experiment on the noisy speech (white noise) utterance /water/ of a female speaker of various input SNRs and obtain the 1st estimated DON1, Z_{lm} and 2nd estimated DON2, Z_{2m} . Figure 2 illustrates the frame wise true degree of noise calculated and the estimated degree of noise obtained in every analysis frame for different input SNRs. By adopting smoothing in the MVS, the overestimation problem is minimized and the effect of *musical noise* is reduced. In fact smoothing is performed to reduce the high frequency fluctuations. Since for speech most of the signal energy is concentrated in low frequencies, for that reason smoothing is reducing the high frequency components and gives increased

Vol. 8, No. 8, November 2010

signal-to-noise ratio. The Fig. 3 shows, the true and the estimated degree of noise are almost equal in all SNRs.

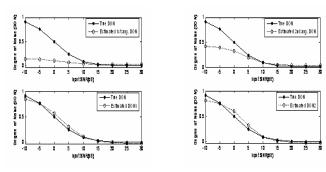


Figure 2a. True vs 1st avg. DON (T) and True vs 1st estimated DON1 (B).

Figure 2b. True vs 2^{nd} avg. DON (T) and True vs 2^{nd} estimated DON2 (B).

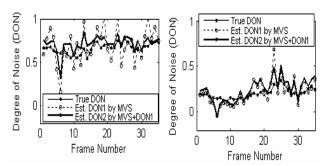


Figure 3. Frame wise graphical representations of the true (solid with point) and the 1^{st} estimated DON1 (dotted line with circle) and 2^{nd} estimated DON2 (solid line with double linewidth) for -5dB (left) and 5dB (right) SNR noisy speech.

C. Estimation of Noise Spectrum

The noise spectrum is estimated from the SMVS and 1st estimated DON according to the condition

$$D_m(k) = Y_{\min}(k) + \left(\sqrt{Z_{1m}} \times Y_{rms}\right)$$
(4)

Then we made some updates of $D_m(k)$, the updated spectrum is again smoothed by three point moving average, and lastly the main maximum of the spectrum is identified and are suppressed [7].

III. WEIGHTED NOISE SUBTRACTION (NWNS)

Noise reduction based on implementation of the traditional spectral subtraction (SS) require an available estimation of the embedded noise, here, in time domain we named Noise Subtraction (NS). It is observed that, in NS, degradation occurs for overestimation of noise within the unvoiced region of noisy speech at higher input SNR (>10 dB). We manually seen that the unvoiced region provides flat spectrum characteristics and exhibits low SNR that gives more degree of noise value that increases the noise level. Therefore, the extracted noise in unvoiced region is high and degrades the speech. From Figure 4, it is seen that the unvoiced frame of higher SNR (>10 dB) input noisy speech provides flat spectrum and low SNR that gives more DON2 (Z_{2m}) value that increases weighting factor. So more noise has subtracted at every unvoiced frame than from every voiced frame, say at 25 dB SNR input speech. Consequently speech distortion has to be occurred. For that, we introduce a nonlinear weighting factor to control the overestimation and minimizing the effect of residual noise. The NWNS is given by:

$$s_{1}(n) = y(n) - \sqrt{\alpha \times Z_{tr} \times \hat{d}_{ss}(n)}$$
(5)

where $\alpha = 0.3019 + 6.4021 \times Z_{2m} - 14.109 \times Z_{2m}^2 + 9.8273 \times Z_{2m}^3$ is nonlinear weighting factor.

It is observed from Eq. (5) that it needs the input SNR. The input SNR can be estimated using variance is given by

$$SNR_{input} = 10 \log_{10} \left(\frac{\sigma_s^2}{\sigma_\eta^2} \right)$$
 (6)

where, σ_s^2 and σ_η^2 are the variances of speech and noise, respectively. We assume that due to the independency of noise and speech, the variance of the noisy speech is equal to the sum of the speech variance and noise variance. It is found that by adopting nonlinear weighted in NS, a good noise reduction is obtained. Although with the NWNS, we find the good performance with less musical noise by informal listening test.

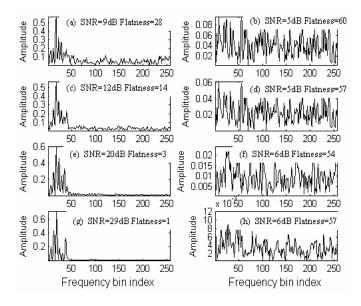


Figure 4. The depictions spectrums of voiced and unvoiced frames degraded by white noise at 5dB SNR is shown in (a) and (b), 10dB SNR is shown (c) and (d), 20dB SNR is shown in (e) and (f), 30dB SNR is shown in (g) and (h) respectively.

A. Derievation of non linear weight

It is observed that the outcome of the subtraction type algorithms produce musical noise and that cannot be avoided. Since algorithms with fixed subtraction parameters are unable to adapt well to the varying noise levels and characteristics, therefore it becomes imperative to estimate a suitable factor to update the noise level. Hence we derive a nonlinear weighting factor α for this purpose. First, simulation is performed over 7 males and 7 females speakers of different sentences at different SNR levels, randomly selected from the TIMIT database, for different values of α and record the output SNR.

Table 1 shows the performance of computer simulation of the algorithm of a given noisy sentence of a female speaker for different values of α .

TABLE 1: The output SNR for a noisy speech of a female speaker for different values of α . for wide range of input SNR (-10dB to 30dB). The speech is degraded by white noise nose.

Input SNR→	-10dB	-5dB	OdB	5dB	10dB	15dB	20dB	25dB	30dB
Values of α	1000		- Tab						3000
0.01	-9.9456	-4.9444	0.05659	5.0557	10.05	15.045	20.043	25.041	30.038
0.02	-9.891	-4.8886	0.11347	5.1117	10.1	15.09	20.085	25.079	30.069
0.03	-9.836	-4.8324	0.17066	5.168	10.151	15.134	20.126	25.116	30.094
0.04	-9.7808	-4.7759	0.22815	5.2244	10.201	15.179	20.167	25.15	30.113
0.05	-9.7252	-4.7191	0.28594	5.2812	10.252	15.223	20.207	25.183	30.126
0.06	-9.6693	-4.662	0.34404	5.3382	10.302	15.268	20.247	25.213	30.132
0.07	-9.6131	-4.6045	0.40245	5.3954	10.353	15.312	20.286	25.242	30.131
0.08	-9.5566	-4.5468	0.46117	5.4529	10.404	15.356	20.324	25.268	30.124
0.09	-9.4998	-4.4887	0.52021	5.5106	10.455	15.4	20.361	25.291	30.11
0.1	-9.4426	-4.4303	0.57955	5.5686	10.506	15.444	20.398	25.313	30.09
0.2	-8.8528	-3.828	1.1907	6.1623	11.021	15.873	20.723	25.395	29.568
0.3	-8.2277	-3.1904	1.835	6.78	11.541	16.277	20.955	25.233	28.615
0.4	-7.5644	-2.5153	2.5132	7.419	12.059	16.643	21.077	24.853	27.465
0.5	-6.8602	-1.8012	3.2245	8.0736	12.564	16.96	21.079	24.307	26.28
0.6	-6.1134	-1.0481	3.9654	8.7342	13.044	17.211	20.961	23.652	25.137
0.7	-5.324	-0.25903	4.7275	9.385	13.482	17.384	20.733	22.941	24.069
8.0	-4.4958	0.55759	5.4948	10.002	13.858	17.47	20.411	22.209	23.082
0.9	-3.6392	1.384	6.2397	10.552	14.151	17.462	20.015	21.481	22.172
1	-2.7765	2.1879	6.919	10.995	14.34	17.362	19.565	20.772	21.333
1.1	-1.9486	2.9164	7.4733	11.286	14.412	17.175	19.081	20.089	20.558
1.2	-1.2218	3.4954	7.8352	11.393	14.359	16.912	18.576	19.435	19.839
1.3	-0.68597	3.8413	7.9487	11.3	14.188	16.587	18.063	18.813	19.17
1.4	-0.43312	3.891	7.7942	11.021	13.91	16.213	17.549	18.221	18.544
1.5	-0.51612	3.6343	7.3982	10.588	13.545	15.805	17.042	17.658	17.958
1.6	0.91692	3.1205	6.82	10.044	13.116	15.373	16.545	17.123	17.406

TABLE 2: The average weight of α for 7 male and 7 female utterances corresponding to wide range of input SNR (-10dB to 30dB).

Input SNR→	-10dB	-5dB	0dB	5dB	10dB	15dB	20dB	25dB	30dB
	-1000	-5425	VIII.	3413	1000	1541	2000	2500	30th
Speaker									
fcjfo	1.4	1.4	1.3	1.2	1.1	0.8	0.5	0.2	0.06
mapvo	1.4	1.3	1.1	0.9	0.6	0.3	0.1	0.03	0.01
fdmlo	1.4	1.4	1.3	1.2	1.1	0.8	0.4	0.2	0.05
makro	1.4	1.3	1.2	1.1	0.7	0.4	0.1	0.05	0.02
fltmo	1.4	1.4	1.3	1.1	0.9	0.6	0.3	0.1	0.03
mtjso	1.4	1.4	1.3	1.2	0.9	0.5	0.2	0.08	0.03
fmjfo	1.4	1.4	1.3	1.3	1.3	1.1	0.7	0.4	0.1
fntbo	1.4	1.3	1.3	1.1	0.9	0.6	0.2	0.09	0.03
mcalo	1.4	1.3	1.2	0.9	0.6	0.3	0.1	0.04	0.01
mdmto	1.4	1.3	1.2	1.0	0.6	0.3	0.1	0.04	0.01
fvmho	1.4	1.4	1.3	1.3	1.1	0.9	0.5	0.2	0.08
mdpko	1.4	1.3	1.2	1.0	0.6	0.3	0.1	0.04	0.01
fpazo	1.4	1.4	1.3	1.2	1.0	0.7	0.3	0.1	0.04
mklwo	1.4	1.3	1.2	1.0	0.7	0.3	0.1	0.05	0.01
Average	1.4	1.35	1.25	1.1071	0.86429	0.56429	0.26429	0.11571	0.035

Let the set of data points (x_i, y_i) , i = 1, 2, ..., 9 and the curve given by Y = f(x) be fitted for this data. At $x = x_i$, the experimental value of the ordinate is y_i and the corresponding value on the fitting curve is $f(x_i)$. If e_i is the error of approximation at $x = x_i$, then $e_i = y_i - f(x_i)$, then the summation of the square of the errors is given by

$$SE = \sum_{i=1}^{9} e_i^2$$

We consider α is a polynomial of degree 3.

Then the 3rd degree polynomials are:

$$\alpha = f(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3$$
 (7)

be fitted to the data points (x_i, y_i) , i = 1, 2, ..., 9.

x represents the values of DON2.

The summation of errors at $x = x_i$ is given by

$$SE = \sum_{i=1}^{9} \left[y_i - \left(a_0 + a_1 x_i + a_2 x_i^2 + a_3 x_i^3 \right) \right]^2$$
 (8)

For SE to be minimum, we have

$$\frac{\partial (SE)}{\partial a_0} = -2\sum_{i=1}^9 \left[y_i - \left(a_0 + a_1 x_i + a_2 x_i^2 + a_3 x_i^3 \right) \right]^2 = 0$$
(9)

$$\frac{\partial (SE)}{\partial a_1} = -2\sum_{i=1}^9 \left[y_i - \left(a_0 + a_1 x_i + a_2 x_i^2 + a_3 x_i^3 \right) \right]^2 x_i = 0$$
 (10)

$$\frac{\partial (SE)}{\partial a_2} = -2\sum_{i=1}^9 \left[y_i - \left(a_0 + a_1 x_i + a_2 x_i^2 + a_3 x_i^3 \right) \right]^2 x_i^2 = 0$$
(11)

$$\frac{\partial (SE)}{\partial a_3} = -2\sum_{i=1}^9 \left[y_i - \left(a_0 + a_1 x_i + a_2 x_i^2 + a_3 x_i^3 \right) \right]^2 x_i^3 = 0$$
 (12)

From Eq. (9), (10), (11) and (12) we have,

$$\sum_{i=1}^{9} y_i = 9a_0 + a_1 \sum_{i=1}^{9} x_i + a_2 \sum_{i=1}^{9} x_i^2 + a_3 \sum_{i=1}^{9} x_i^3$$
(13)

$$\sum_{i=1}^{9} x_i y_i = a_0 \sum_{i=1}^{9} x_i + a_1 \sum_{i=1}^{9} x_i^2 + a_2 \sum_{i=1}^{9} x_i^3 + a_3 \sum_{i=1}^{9} x_i^4$$
(14)

$$\sum_{i=1}^{9} x_i^2 y_i = a_0 \sum_{i=1}^{9} x_i^2 + a_1 \sum_{i=1}^{9} x_i^3 + a_2 \sum_{i=1}^{9} x_i^4 + a_3 \sum_{i=1}^{9} x_i^5$$
(15)

$$\sum_{i=1}^{9} x_i^3 y_i = a_0 \sum_{i=1}^{9} x_i^3 + a_1 \sum_{i=1}^{9} x_i^4 + a_2 \sum_{i=1}^{9} x_i^5 + a_3 \sum_{i=1}^{9} x_i^6$$
(16)

We write these equations in a matrix form as:

$$\begin{bmatrix} \sum_{i=1}^{9} y_i \\ \sum_{i=1}^{9} x_i y_i \\ \sum_{i=1}^{9} x_i^2 y_i \\ \sum_{i=1}^{9} x_i^2 y_i \end{bmatrix} = \begin{bmatrix} 9 & \sum_{i=1}^{9} x_i & \sum_{i=1}^{9} x_i^2 & \sum_{i=1}^{9} x_i^2 \\ \sum_{i=1}^{9} x_i & \sum_{i=1}^{9} x_i^2 & \sum_{i=1}^{9} x_i^3 & \sum_{i=1}^{9} x_i^4 \\ \sum_{i=1}^{9} x_i^2 & \sum_{i=1}^{9} x_i^3 & \sum_{i=1}^{9} x_i^4 & \sum_{i=1}^{9} x_i^5 \\ \sum_{i=1}^{9} x_i^3 & \sum_{i=1}^{9} x_i^4 & \sum_{i=1}^{9} x_i^5 & \sum_{i=1}^{9} x_i^6 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix}$$

$$(17)$$

Eq.(17) is a Vander monde matrix. We can also obtain the matrix for a least squares fit by writing:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y \\ y_7 \\ y_8 \\ y_9 \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 \\ 1 & x_2 & x_2^2 & x_2^3 \\ 1 & x_3 & x_3^2 & x_3^3 \\ 1 & x_4 & x_4^2 & x_4^3 \\ 1 & x_5 & x_5^2 & x_5^3 \\ 1 & x_6 & x_6^2 & x_6^3 \\ 1 & x_7 & x_7^2 & x_7^3 \\ 1 & x_8 & x_8^2 & x_8^3 \\ 1 & x_9 & x_9^2 & x_9^3 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ a_3 \end{bmatrix}$$

$$(18)$$

In matrix notation, Eq.(18) can be written as:

$$Y = XA \tag{19}$$

where.

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \end{bmatrix} \quad \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 \\ 1 & x_2 & x_2^2 & x_2^3 \\ 1 & x_3 & x_3^2 & x_3^3 \\ 1 & x_4 & x_4^2 & x_4^3 \\ 1 & x_5 & x_5^2 & x_5^3 \\ 1 & x_6 & x_6^2 & x_6^3 \\ 1 & x_7 & x_7^2 & x_7^7 \\ 1 & x_8 & x_8^2 & x_8^3 \\ 1 & x_9 & x_9^2 & x_9^3 \end{bmatrix} \quad and \quad A = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix}$$

Multiply both sides of Eq.(19) by X^T (transpose of X)

$$X^T Y = X^T X A (20)$$

This matrix equation can be solved numerically, or can be inverted directly it is well formed, to yield the solution vector

$$A = \left(X^T X\right)^{-1} X^T Y \tag{21}$$

In our experiment,

 $\begin{aligned} x_i &= DON\,2_i = \begin{bmatrix} 0.80707, 0.75235, 0.59967, 0.32374, 0.095033, 0.01379, -0.009902, -0.01741, -0.019616 \end{bmatrix} \\ y_i &= \alpha_i = \begin{bmatrix} 1.4, & 1.35, & 1.25, & 1.1071, & 0.86429, & 0.56429, & 0.26429, & 0.11571, & 0.035 \end{bmatrix} \end{aligned}$

So, we have

$$X = \begin{bmatrix} 1 & DON2_1 & DON2_1^2 & DON2_1^3 \\ 1 & DON2_2 & DON2_2^2 & DON2_2^3 \\ 1 & DON2_3 & DON2_3^2 & DON2_3^3 \\ 1 & DON2_4 & DON2_4^2 & DON2_3^3 \\ 1 & DON2_5 & DON2_5^2 & DON2_5^5 \\ 1 & DON2_6 & DON2_6^2 & DON2_6^3 \\ 1 & DON2_7 & DON2_7^2 & DON2_7^7 \\ 1 & DON2_8 & DON2_8^2 & DON2_8^3 \\ 1 & DON2_9 & DON2_9^2 & DON2_9^3 \end{bmatrix}$$
 and $Y = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \\ \alpha_6 \\ \alpha_7 \\ \alpha_8 \\ \alpha_9 \end{bmatrix}$

Finally we put the values of DON2₁,....DON2₉ to get X and put the value of $\alpha_1,....,\alpha_9$ to get Y. Therefore, from Eq.(21), we have

$$A = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 0.3019 \\ 6.4021 \\ -14.109 \\ 9.8273 \end{bmatrix}$$

Now substitute the value of a_0, a_1, a_2 and a_3 in Eq.(7)

$$\alpha = 0.3019 + 6.4021 \times DON2 - 14.109 \times DON2^{2} + 9.8273 \times DON2^{3}$$
(22)

Equation (22) is the derivation of the nonlinear weighting factor α and is used in Eq. 5.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

The proposed noise estimation method is compared with the conventional noise estimation algorithm using MVS in terms of noise estimation accuracy and quality. Figures 5 illustrate results of noise estimation in frequency domain (FD) measure. In the experiment, we consider the vowel phoneme sound /oy/, degraded by the white noise at 0dB SNR. It shows that, by adopting the proposed DON1 (Z_{1m}), it is possible to estimate the state of the added noise more precisely. We achieve sufficient improvements in noise amplitudes using the MVS+DON1 estimator. Objective measure is also performed to verify the quality of the estimated noise. For that we use the PESQ MOS measure. Figure 6 shows the PESQ MOS value between the added and the estimated noise at different noise levels. It shows that PESQ MOS value gradually decreases at the higher SNR.

To study the speech enhancement performance, an experiment is carried out by taking 56320 samples of the clean speech /she had your dark suit in greasy wash water all year/from TIMIT database. The speech signal is corrupted by white, pink and HF channel noises at various SNR levels are taken from NOISEX database. The results of the average output SNR obtained from for white noise, pink noise and HF channel noise at various SNR levels are given in Table 1 for NS and NWNS, respectively.

Vol. 8, No. 8, November 2010

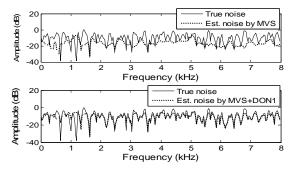


Figure 5. Noise spectrums (original and estimated).

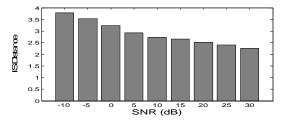


Figure 6. Estimated noise quality based on PESQ MOS.

We observe from the Tables 3 that the overall output SNR by NS is improved up to 10 dB input SNR and degraded from 15 dB and higher. Degradation occurs for overestimation of noise within the unvoiced region of noisy speech at higher input SNR (>10 dB). Since the unvoiced region provides flat spectrum characteristics and exhibits low SNR gives more DON2 value that increases the noise level. Consequently the extracted noise in unvoiced region is high that is responsible to degrade the speech. Hence it is essential to add a weighting factor to control the overestimation and we have a better performance by NWNS throughout the SNR. It is observed that the enhanced speech is distorted in low voiced parts due to remove the noise in NS method whereas NWNS does not. But little amount of noise can be removed from the corrupted speech by NWNS method. So in NS method there is a loss of speech intelligibility while NWNS maintains it. We have found better results compared to our previous study [7] for a wide range of SNRs.

TABLE 3: The results of average output SNR for various types of noise at different input SNR by the NS and NWNS methods.

Input	White	White noise		nel noise	Pink no	Pink noise	
SNR	NS	NWNS	NS	NWNS	NS	NWNS	
-10dB	-2.8	-1.57	-7.4	-7.5	-7.1	-7.1	
-5dB	2.0	2.4	-2.3	-2.7	-2.2	-2.3	
0dB	6.5	5.3	2.6	1.9	2.6	2.2	
5dB	10.3	8.7	7.3	6.4	7.3	6.4	
10dB	13.3	11.7	11.5	10.8	11.3	10.8	
15dB	15.4	15.8	14.5	15.4	14.4	15.4	
20dB	16.7	20.4	16.4	20.2	16.3	20.3	
25dB	17.5	25.2	17.3	25.1	17.3	25.2	
30dB	17.7	30.1	17.7	30.1	17.6	30.1	

CONCLUSIONS

In this paper, an improved noise estimation technique is discussed. Initially noise is estimated from the valleys of the amplitude spectrum. Then we have adjusted the estimated noise amplitudes by the estimated DON1. It eliminates the need for a VAD by exploiting the short time characteristics of speech signals. In the result part, it is shown that the state of the added noise is more accurate with MVS+DON1. The enhanced speech using time domain nonlinear weighted noise subtraction results in sufficient noise reduction. The main advantage of the algorithm is the effective removal of the noise components for a wide range of SNRs. We not only have better SNR but also a better speech quality with significantly reduced residual noise. However, a little noisy effect still remains. This issue will be addressed in our future study.

REFERENCES

- [1] Benesty, J., Makino, S., and Chen, J., Speech Enhancement, Springer-Verlag Berlin Heidelberg, 2005.
- Martin, R., and Lotter, T., "Optimal Recursive Smoothing of Non-Stationary Periodograms", Proc. IWAENC, pp. 167-170, Sept. 2001.
- [3] Cohen, I., and Berdugo, B., "Noise Estimation by Minima Controlled Recursive Averaging for Robust Speech Enhancement", IEEE Signal Processing Letters, vol. 9, no. 1, pp. 12-15, Jan. 2002.
- Cohen, I., "Noise Estimation in Adverse Environments: Improved Minima Controlled Recursive Averaging", IEEE Trans. on Speech and Audio Process., vol. 11, pp. 466-475, Sept. 2003..
- Doblinger, G., "Computationally Efficient Speech Enhancement by Spectral Minima Tracking in Subbands", Proc. EUROSPEECH, pp. 1513-1516, 1995.
- Hirsch, H. G., and Ehrlicher, C., "Noise Estimation Methods for Robust Speech Recognition", Proc. ICASSP, pp. 153-156, 1995.
- Hamid, M. E., Ogawa, K., and Fukabayashi, T., "Noise estimation for Speech Enhancement by the Estimated Degree of Noise without Voice Activity Detection", Proc. SIP 2006, pp. 420-424, Hawaii, August 2006.
- [8] Martin, R., "Speech enhancement using MMSE short time spectral estimation with Gamma distributed speech priors", in Proc. Int. Conf. Speech, Acoustics, Signal Processing, vol. I, pp. 253-256, 2002.
- Martin, R., "Spectral Subtraction Based on Minimum Statistics", Proc. EUSIPCO, pp. 1182-1185, 1994.
- [10] Martin, R., "Statistical Methods for the Enhancement of Noisy Speech", Proc. IWAENC2003, pp. 1-6, 2003.

AUTHORS PROFILE



Md. Ekramul Hamid received his B.Sc and M.Sc degree from the Department of Applied Physics and Electronics, Rajshahi University, Bangladesh. After that he obtained the Masters of Computer Science degree from Pune

University, India. He received his PhD degree from Shizuoka University, Japan. During 1997-2000, he was a lecturer in the Department of Computer Science and Technology, Rajshahi University. Since 2007, he has been serving as an associate professor in the same department. He is currently working as an assistant professor in the college of computer science at King Khalid University, Abha, KSA. His research interests include speech enhancement, and speech signal processing.



Md. Zasim Uddin received his Bsc and MSc in Computer Science & Engineering from Rajshahi University, Rajshahi, Bangladesh. He has been awarded National Science and Information & Communication Technology

Fellowship (Government of the People's Republic of Bangladesh) in 2009. Currently he is a lecturer of Computer Science & Engineering department, Dhaka International University, Dhaka, Bangladesh. His research interests include medical image and signal processing. He is a member of Bangladesh Computer Society.



Md. Humayun Kabir Biswas, working as an international lecturer in the Department of Computer Science at King Khalid University, Kingdom of Saudi Arabia. Before joining at KKU, he worked as a lecturer under the

Department of Computer Science and Engineering at IUBAT-International University of Business Agriculture and Technology, Bangladesh. He has completed his Master of Science in Information Technology degree from Shinawatra University, Bangkok, Thailand. He is keen to doing research on semantic web, intelligent information retrieval technique and Machine Learning Technique. His current research interest is audio and image signal processing.

Somlal Das received B.Sc (Hons) and M.Sc. degrees from the Department of Applied Physics and Electronics in the University of Rajshahi, Bangladesh. He joined as a lecturer at the Department of Computer Science and Engineering in the University of Rajshahi, Bangladesh, in 1998. He is currently serving as an Assistant Professor and working as Ph.D. student at the same Department. His research interests are in speech signal processing, speech enhancement, speech analysis, and digital signal processing.

Scalable Video Coding in Online Video transmission with Bandwidth Limitation

¹Sima Ahmadpour, ²Salah Noori Saleh, ¹Omar Amer Abouabdalla, ¹Mahmoud Baklizi, ¹Nibras Abdullah

1: National Advanced IPv6 Center of Excellence 1: Universiti Sains Malaysia 1: Penang, Malaysia

Abstract— Resource limitation and variety of network and users cause many obstacles while transmitting data especially online video data through network. Video applications in Internet face by significant growth in several market segments and bandwidth limitation is one of those challenges which consider as a main obstacle in this paper.

 ${\it Keywords-component; bandwidth \ limitation, \ video \ codec, \ video \ conferencing, \ SVC}$

I. Introduction

During the last few years the Internet has grown tremendously and has penetrated all aspects of everyday life. Therefore, people are willing to communicate and exchange the information due to update their knowledge in different aspects. In this case, live connection among people in different location around the world comes to demand. Interactive video services like video conferencing based on online education, distance learning, online video games over the Internet are mostly popular now a days. Meanwhile, modern video transmission works mainly based on Real Time Transport Protocol (RTP/IP) for real time services [1]. RTP mainly is an Internet protocol to transmit real time multimedia data over either unicast or multicast network services.

Typically, video data need to be compressed and decoded by video CODECs in the real time applications. CODECs are one of the common solutions to adapt video streams with low bandwidth over the network. Bandwidth limitation is a main obstacle faces by video streams through network especially in heterogeneous networks. [4], [8], [10], [12], [9]. In fact, video CODECs are computer programs for compressing video due to reduce bandwidth requirements. After that, they transmit those compressed streams and decrease the storage requirements to archive them easily. Solving the existed mismatches between bandwidth and computational requirement help to identify the minimum channel bandwidth required to pass encoded stream and minimize the specification of decoding device. In general, the video is encoded with CODEC once as a stream. Later, the resulting stream would produce a full resolution video in decode step [4].

This paper focuses on the bandwidth limitation of internet through transmitting online video data as the main concern. Not only bandwidth limitation on current Internet but also, extreme bandwidth requirement of video lead to have a proper resource management to reach the real time performance. In section II five existing CODEC standards are defined. Comparison points and their performance through transmitting online video streams are noted in a table at section III. Finally, the conclusion has been presented.

II. CODEC STANDARDS

A codec is whether a program or a device to encode or decode a digital data stream or a signal. Normally, a codec use to encode the streams for transmission, storage or encryption. On the other hand, a stream may be decoded for playback or editing. The major video coding standards have been created since 1990s. They are mostly formulated based on the same generic design of a video CODEC that combines a motion estimation and compensation front end, a transform stage and an entropy encoder. Generally, each standard explains two terms such as:

- Compressed form visual data known as coded syntax.
- The method of decoding the syntax to rearrange the visual information.

The main target of each standard is to make sure both encoders and decoders are compatible to work properly with each other. Otherwise, suppliers are not able to develop proper products [2].

There are different types of CODECs in the market since 1990. They got huge changes facing by new production regarding to their new requirements each time. In this paper MPEG-2, H.263, MPEG-4 Visual, H.264 and SVC are discussed.

A. MPEG-2

MPEG-2 mostly is used for broadcasting digital TV via cable, DVD-Video and MPEG Layer 3 audio coding known as MP3 which is became popular in terms of storage and sharing music. MPEG-2 introduced the idea of Profiles and Levels for the first time without restricting the flexibility of the standard. In addition, MPEG-2 applications were looking for the proper standard for the next generation of products [2].

B. H.263

The original H.263 standard has published as a standard in 1995. Its powerful compression was supporting basic video quality at bitrates of below 30 kbit/s. H.263 is compatible with the standards over wide range of circuit and packet switched networks. H.263 contains four optional coding modes and added some extra modes to support improved compression efficiency and robust transmission over lossy networks [2].

C. MPEG-4 Visual

MPEG-4 is Part2 of the MPEG-4 group of standards. It is developed by Moving Picture Experts Group known as (MPEG). It works on coding the audio-visual objects and can support those applications which are noted as below:

- Legacy video applications like TV broadcasting, video conferencing and video storage.
- Object-based video applications that a video scene contains a combination of different distinct video which are coded independently.
- Computer graphics which are using 2D and 3D deformable mesh geometry or human faces and bodies that are animated.
- Hybrid video applications which are combining natural video, images and graphics generated by computer.
- Streaming video over the Internet and mobile channels.
- High-quality video editing using for the studio production.

In overall, MPEG-4 Visual is slightly simple video coding mechanism with block-based video CODEC using motion compensation followed by DCT, quantization and entropy coding [2].

D. H.264 codec

H.264 is known as Advanced Video Coding which is a standard for codec visual data. It is actually designed to support

a strong and efficient coding and transmit of rectangular video frames. H.264 was created to improve the functionality of previous CODEC such as H.263 and MPEG-4.

H.264 mostly uses to compress video in both commercial and military application. H.264 works based on MPEG-2, using macroblock-based motion prediction and allows more flexible encoding rather than MPEG-2. Furthermore, H.264 explains just bit stream syntax. Considering both error resilience and coding efficiency, different H.264 encoders may create different output [10]. The main applications contain:

- Two-way video communication such as video conferencing or video telephony.
- Coding for broadcast and high quality video.
- Video streaming over packet networks [2].

E. Scalable Video Codec

The Scalable Video Coding is an extension of H.264/AVC which is created to control bandwidth and loss resilient video streaming. SVC is working as a multilayer predictive encoder. Therefore, users are able to adapt the received videos by extracting and decoding the code layers based on the capability of their own devices and network throughput [1], [11].

In Scalable Video Coding, decoder is able to decode just part of the decoded bitstream selectively. Apart from that, encoder arranges the coded stream based on layers including a base layer and one or more enhancement layers. For instance in (figure1), decoder A can decode a basic quality version of video scene which is received from only the base layer. On the other hand, decoder B is receiving all layers and decodes a high quality version of the video. Furthermore, a number of applications come to demand for example, a low complexity decoder may able to decode the base layer; a low bitstream with restricted capacity may be extracted for transmission over a network segment and also an error sensitive base layer may be transmitted with higher priority in comparison with enhancement layers.

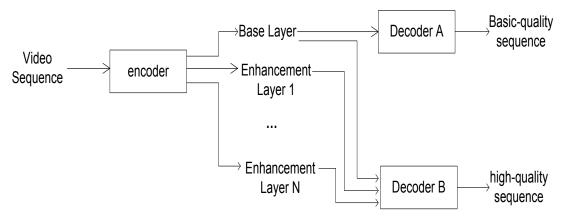


Figure 1. Scalable coding: general concept [2]

The Scalable Video CODEC influences the tools and concepts of original CODEC. In addition, SVC supports a number of scalable coding modes such as Spatial Scalability, Temporal

Scalability, quality scalability as well as Fine Grain Scalability [1], [2], [4], [11], [12], [13]. It can cause different decoded video due to different frame rate, resolution or different quality.

Vol. 8, No. 8, 2010

1) Spatial Scalability

Spatial Scalability is related to resolution of each frame. Typically, the base layer includes a reduced resolution of each coded frame. The output of base layer encoding is a low resolution sequence while encoding both base layer and enhancement layers creates a high resolution output. Meanwhile, spatial enhancement layers is coding deference information among the picture of up scaled reference layer and higher resolution of original picture increase a resolution of coded pictures [2]

2) Temporal Scalability

Temporal scalability can be used to create a high frame rate. Generally, it is related to the number of frames per second (fps) in video stream. Temporal enhancement layers encode additional pictures to increase the frame rate of reproduced video. However, it is not allowed to use those pictures as reference for spatial enhancement layers [3]

3) Quality Scalability

With Quality Scalability, the substream creates a complete bitstream with a spatio-temporal resolution and lower fidelity. Quality Scalability is often related to signal to noise Ratio (SNR) and it is mainly related to visual quality layers of the coded video by different bitrates [3].

4) Fine Grain Scalability

Fine Grain Scalability (FGS) increases the quality of sequence in small steps. A FGS application streams video through network that may be suitable to scale coded stream being match with available bit rate [2]. The main new features of SVC are as follow:

- Variable block-size motion-compensated prediction with the block size down to 4x4 pixels;
- Quarter-pixel motion vector accuracy;
- Multiple reference pictures for motion compensation;
- Bi-directional predicted picture as a reference for motion prediction;
- Intra-picture prediction in the spatial domain;
- Adaptive deblocking filter within the motioncompensated prediction loop;
- Small block-size transformation (4x4 block transform);
- Enhanced entropy coding methods: Context-Adaptive Variable-Length Coding (CAVLC) and Context-Adaptive Binary Arithmetic Coding (CABAC)).

	MPEG-2	H.263	MPEG-4 Visual	H.264	SVC
Data type	Object base coding scheme	Object base wide range of circuit and packet switched networks	Rectangular video frames and fields	Rectangular video frames and fields	Rectangular video frames and fields
Compression efficiency	Loss efficiency	Loss efficiency	Medium	High	High
Motion compensation Minimum block size	8 × 8	8 × 8	8 × 8	4 × 4	4 × 4

TABLE I. COMPARISON OF ABOVE MENTIONED CODECS

III. DISCUSSION

CODEC's improvements increase extremely along with multimedia developments. Considering above mentioned CODECs' capabilities, each CODEC has been adapted by the newest version of production at its own time. Therefore, MPEG-2 and H.263 stand on object base scheme where loss Compression efficiency is another issue.

In MPEG-4 Visual and H.264 macroblock-based motion compensation is the core technology of video coding while transformation and quantization of residual data is considered [2]. Although, the compression efficiency of MPEG-4 Visual is medium in comparison with H.264 and SVC, still it works better than both H.263 and MPEG-2 with loss efficiency.

SVC is able to create high compression efficiency coding which is the main requirement of online applications. In case, when a lower resolution or bandwidth stream is needed to reach the network and a lower performance device is aimed a small part of the decoded stream would be sent without any

further processing. Current small stream would be easier to decode and result a video with low resolution. In this case, the encoded video stream would adapt itself to the bandwidth of the transport channel and to requirements of target device. This is the characteristic of Scalable Video CODEC.

IV. CONCLUSION

Although traditional systems for transmitting video data may have some scalable capability, but they still have some challenges due to less coding efficiency and complex decoder. Furthermore, different types of scalability may be combined and create large number of representations with different spatio- temporal resolutions. In this case, one single bit stream is able to support the bit rates. However, SVC supports is using multiple dimension scalable modes to support flexible bitstream. Otherwise, the main problem of SVC is that: how to guarantee user perceived quality (UPQ) which means that how to guarantee user satisfaction with video quality in video services. This is considered as future work [7].

REFERENCES

- H. Schwarz, D. Marpe, and T. Wiegand, Overview of the Scalable Video Coding Extension of the H.264/AVC Standard, IEEE, 2007
- [2] I. E. G. Richardson, H.264 and MPEG-4 Video Compression, e-book, 2003
- [3] Z. He, Y. Yan, and Y. perieto, Temporal Scalability for Low Delay Scalable Video Coding, US 2009/0060035 A1, Mar.5, 2009
- [4] The Stretch Scalable Video CODEC, Extending the Possibilities >www.stretchinc.com
- [5] S. Rimac-Drlje, O. Nemþiü, and M. Vranješ, Scalable Video Coding Extension of the H.264/AVC Standard, International Symposium ELMAR-2008, 10-12 September 2008, Zadar, Croatia
- [6] J. Ohm, Introduction to MPEG-2 Video (13818-2), International Organization for Standardization Organization International Normalisation.
- [7] W. Song, D. Tjondronegoro, and S. Azad, User-Centered Video Quality Assessment for Scalable Video Coding of H.264/AVC Standard, SpringerLink 2008

- [8] M. Sadegh Talebi, A. Khonsari, and M. H. Hajiesmaili, Utility-proportional bandwidth sharing for multimedia transmission supporting scalable video coding, ELSEVIER 2009.
- [9] F. Verdicchio, Y. Andreopoulos, T. Clerckx, J. Barbarien, A. Munteanu, J. Cornelis, and P. Schelkens, Scalable Video Coding Based on Motion-Compensated Temporal Filtering: Complexity and Functionality Analysis.
- [10] M. Brown, D. Bushmitch, K. Kerpez, D. Waring, and Y. Wang, Low-Bit Rate Video Codec Parameter Evaluation and Optimization, IEEE 2009.
- [11] C. M. Lin, J. K. Zao, W. H. Peng, C. C. Hu, H. M. Chen, and C. K. Yang, Bandwidth Efficient Video Streaming Based Upon Multipath SVC Multicasting, IEEE 2008.
- [12] H. L. Cycon, D. Marpe, T. C. Schmidt, M. W"ahlisch, and M. Winken, Optimized Temporal Scalability for H.264 based Codecs and its Applications to Video Conferencing, IEEE 2010.
- [13] G. Nur, H. Kodikara Arachchi, S. Dogan, and A. M. Kondoz, EVALUATION OF QUALITY SCALABILITY LAYER SELECTION FOR BIT RATE ADAPTATION OF SCALABLE VIDEO CONTENT, IEEE 2009.

Off-Line Handwritten Signature Retrieval using Curvelet Transforms

M. S. Shirdhonkar
Dept. of Computer Science and Engineering,
B.L.D.E.A's College of Engineering and Technology
Bijapur, India

Manesh Kokare
Dept. of Electronics and Telecommunication,
S.G.G.S Institute of Engineering and Technology
Nanded, India

Abstract—— In this paper, a new method for offline handwritten signature retrieval is based on curvelet transform is proposed. Many applications in image processing require similarity retrieval of an image from a large collection of images. In such cases, image indexing becomes important for efficient organization and retrieval of images. These papers address this issue in the context of a database of handwritten signature images and describes a system for similarity retrieval. The proposed system uses a curvelet based texture features extraction. The performance of the system has been tested with an image database of 180 signatures. The results obtained indicate that the proposed system is able to identify signatures with great with accuracy even when a part of a signature is missing.

Keywords- Handwritten recognition, Image indexing, Similarity retrieval, Signature verification, Signature identification.

I. Introduction (Heading 1)

A. Motivation

A signature appears on many types of documents such as bank cheques in daily life and credit slips, thus signature has a great importance in a person's life. Automatic bank cheque processing is an active topic in the field of document analysis and processing. Signature validity confirmation of different document is one of the important problems in automatic document processing. Now a days, person identification and verification are very important in security and resource access control. For this purpose the first and simple way is to use Personal Identification Number (PIN), but PIN code may be forgotten. Now an interesting method to identification and verification is biometric approach [1]. Biometric is a measure for identification that is unique for each person. Always biometric is together with person and cannot be forgotten. In addition biometric usually cannot be misused.

Handwritten signature retrieval is still a challenging work in the situations of a large database. Unlike fingerprint palm print and iris, signatures have significant amount of intra class variations making the research even more compelling. This approach with the potential applications of signature recognition/verification system optimized with efficient signature retrieval mechanism.

B. Related works.

Signature verification contain two areas: off-line signature verification ,where signature samples are scanned into image representation and on-line signature verification, where signature samples are collected from a digitizing tablet which is capable of pen movements during the writing. In our work, we survey the offline signature identification and retrieval. In 2009, Ghandali and Moghaddam have proposed an off-line Persians signature identification and verification based on Image registration, DWT (Discrete Wavelet Transform) and fusion. They used DWT for features extraction and Euclidean distance for comparing features. It is language dependent method [1]. In 2008, Larkins and Mayo have introduced a person dependent off-line signature verification method that is based on Adaptive Feature Threshold (AFT) [2]. AFT enhances the method of converting a simple feature of signature to binary feature vector to improve its representative similarity with training signatures. They have used combination of spatial pyramid and equimass sampling grids to improve representation of a signature based on gradient direction. In classification phase, they used DWT and graph matching methods. In another work, Ramachandra et al [3], have proposed cross-validation for graph matching based offline signature verification (CSMOSV) algorithm in which graph matching compares signatures and the Euclidean distance measures the dissimilarity between signatures.

In 2007, Kovari et. al presented an approach for off-line signature verification, which was able to preserve and take usage of semantic information[4]. They used position and direction of endpoints in features extraction phase. Porwik [5] introduced a three stages method for offline signature recognition. In this approach the hough transform ,center of gravity and horizontal-vertical signature histogram have been employed, using both static and dynamic features that were processed by DWT has been addressed in [6]. The verification phase of this method is based on fuzzy net using the enhanced version of the MDF(Modified Direction feature)extractor has been presented by Armand et.al [7]. The different neural classifier such as Resilient Back Propagation(RBP), Neural network and Radial Basis Function(RBF) network have been used in verification phase of this method. In 1995, Han and Sethi [8], described offline signature retrieval and use a set of geometrical and topological features to map a signature onto

2D strings. We have proposed an offline signature retrieval model based on global features.

The main contribution of this paper is that, we have proposed off-line handwritten signature retrieval using curvelet transform, In retrieval phases Canberra distance measure is used. The experimental results of proposed method were satisfactory and found that it had better results compare with related works. The rest of paper is organized as follows: In section II, discusses the feature extraction phase. The signature retrieval is presented in section III. In section IV, the experimental results and finally section V concludes the work.

II. FEATURE EXTRACTION PHASE

The major task of feature extraction is to reduce image data to much smaller amount of data which represents the important characteristic of the image. In signature retrieval, edge information is very important in characterizing signature properties. Therefore we proposed to use the curvelet transform. The performance of the system is compared with standard discrete wavelet transform which captures information in only three directions.

A. Discrete Wavelet Transform

The multi resolution wavelet transform decomposes a signal into low pass and high pass information. The low pass information represents a smoothed version and the main body of the original data. The high pass information represents data of sharper variations and details. Discrete Wavelet Transform decomposes the image into four sub-images when one level of decomposing is used. One of these sub-images is a smoothed version of the original image corresponding to the low pass information and the other three ones are high pass information that represents the horizontal, vertical and diagonal edges of the image respectively. When two images are similar, their difference would be existed in high-frequency information. A DWT with N decomposition levels has 3N+1 frequency bands with 3N high-frequency bands [9], [10]. The impulse responses associated with 2-D discrete wavelet transform are illustrated in Fig. 1 as gray-scale image.

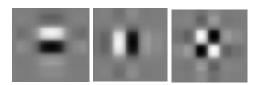


Fig. 1.Impulse response of 0° , 90° and $\pm 45^{\circ}$ of DWT

B. Curvelet Transform

Recently, Candµes and Donoho developed a new multiscale transform which they called the curvelet transform. Motivated by the needs of image analysis, it was nevertheless first proposed in the context of objects f(x1, x2) defined on the continuum plane $(x1, x2) \in \mathbb{R}^2$.

The transform was designed to represent edges and other singularities along curves much more efficiently than traditional transforms, i.e. using many fewer coefficients for a given accuracy of reconstruction. Roughly speaking, to

represent an edge to squared error 1/N requires 1/N wavelets. The curvelet transform, like the wavelet transform, is a multiscale transform, with frame elements indexed by scale and location parameters. Unlike the wavelet transform, it has directional parameters, and the curvelet pyramid contains elements with a very high degree of directional specificity. In addition, the curvelet transform is based on a certain anisotropic scaling principle which is quite different from the isotropic scaling of wavelets. The elements obey a special scaling law, where the length of the support of a frame elements and the width of the support are linked by the relation width ≈ length².see details in [11].

C. Feature Database Creation

To construct the feature vectors of each handwritten signature in the database using DWT and curvelet transform respectively. The Energy and Standard Deviation (STD) were computed separately on each sub band and the feature vector was formed using these two parameter values. The Energy E_k and Standard Deviation σ_k of \mathbf{k}^{th} sub band is computed as follows

$$E_k = \frac{1}{M \times N} \sum_{i=1}^{M} \sum_{i=1}^{N} |W_k(i,j)| \tag{1}$$

$$\sigma_{k} = \left[\frac{1}{M \times N} \sum_{i=1}^{N} \sum_{j=1}^{M} (W_{k}(i, j) - \mu_{k})^{2}\right]^{\frac{1}{2}}$$
(2)

Where $W_k(i,j)$ is the k^{th} wavelet-decomposed sub band, $M\!\!c\!N$ is the size of wavelet decomposed sub band, and μ_k is the mean of the k^{th} sub band. The resulting feature vector using energy and standard deviation are $\bar{f}_E = \begin{bmatrix} E_1 & E_2 & \dots & E_n \end{bmatrix}$ and $\bar{f}_\sigma = \begin{bmatrix} \sigma_1 & \sigma_2 & \dots & \sigma_n \end{bmatrix}$ respectively. So combined feature vector is $\bar{f}_{\sigma \ell} = [\sigma_1 & \sigma_2 & \dots & \sigma_n & E_1 & E_2 & \dots & E_n]$ (3)

III. OFFLINE HANDWRITTEN SIGNATURE RETRIEVAL PHASE

There are several ways to work out the distance between two points in multidimensional space. The most commonly used is the Canberra distance measure. It can be considered the shortest distance between two points. We have used Canberra distance metric as similarity measure. If x and y are the feature vectors of the database and query signature, respectively, x and y have dimension d, then the Canberra distance is given by

Canb (x, y) =
$$\sum_{i=1}^{d} \frac{|x_i - y_i|}{|x_i| + |y_i|}$$

Algorithm 1: Offline Handwritten Signature Retrieval

Input: Test signature: St

Feature database: FV

Output: Distance vector: Dist

Handwritten signature retrieval

Begin

Calculate feature vector of test signature using

DWT and curvelet transform

For each fv in FV do

Dist= Calculate distance between test signature

and fv using (4)

sort Dist

End for

Display the top signature from dist vector.

End

IV. EXPERIMENTAL RESULTS

A. Image Database

The signatures were collected using either black or blue ink (No pen brands were taken into consideration), on a white A4 sheet of paper, with eight signature per page. A scanner subsequently digitized the eight signatures, contained on each page, with a resolution in 256 grey levels. Afterwards the images were cut and pasted in rectangular areas of size 256x256 pixels. Sample signature database for 16 persons are shown in Fig.2. A group of 16 persons are selected for 12 specimen signatures which make the total of 16x12=192 signature database.

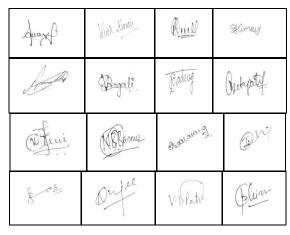


Fig.2. Sample Signature Images Database

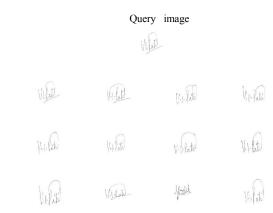


Fig.2. Sample Handwritten Signature Images Database

B. Retrieval Performance

For each experiment, one image was selected at random as the query image from each writer and thus retrieved images were obtained. For performance evaluation of the signature image retrieval system, it is significant to define a suitable metric. Two metrics are employed in our experiments as follows.

$$Recall = \frac{Number \quad of \quad relevant \quad signatures \quad retrieved}{Number \quad of \quad relevant \quad signatures}$$
(5)

$$Precision = \frac{Number \quad of \quad relevant \quad signatures \quad retrieved}{Number \quad of \quad signatures \quad retrieved}$$
(6)

Results correspond to precision and recall rate for a Top1, Top 2, Top 5, Top 8, Top 10, and Top 12. The comparative retrieval performance of the proposed system is shown in Table 1.

Table1: Average Retrieval Performance

	Discrete	wavelet	Curvelet Transform		
	Transform	ı			
Number of	Precision	Recall	Precision	Recall	
Top matches	%	%	%	%	
Top 1	100	8	100	8	
Top 2	80	12.6	96.6	15.4	
Top 5	66.7	28.9	92	36.7	
Top 8	55.8	37.2	73.3	48.5	
Top 10	51.3	43.4	70.7	59.0	
Top 12	47.8	47.5	66.04	65.2	

Retrieval performance of the proposed method is compared using DWT transform technique. We evaluated the performance in terms of average rate of retrieving images as function of the number of top retrieved images. Fig.3 shows graph illustrating this comparison between DWT and curvelet transform according to the number of top matches considered for database. From Fig. 3, it is clear that the new method is superior to DWT. To retrieve images from the database those have a similar writing style to the original request. In Fig. 4, retrieval example results are presented in a list of images having a query image.

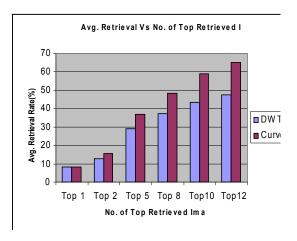


Fig.3. Comparative average retrieval rate using DWT and Curvelet transform

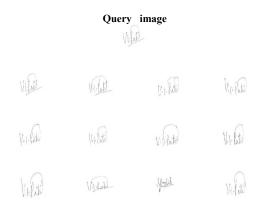


Fig. 4.Sample handwritten signature retrieval example

V. CONCLUSION

Experimental were conducted for quick for retrieval of offline signature and result are presented. The retrieval performance of the proposed method based on edge correspondence is compared with the retrieval method based on DWT. The proposed method is simple, efficient and outperforms the retrieval system based on curvelet features respect to all parameters (Precision, Recall and Correct retrieval). The proposed approach used curvelet features for extracting details and Canberra distance for comparing features.

References

- [1] Samanesh Ghandali and Mohsen Ebrahimi Moghaddam, "Off-Line Persian Signature Identification and Verification based on Image Registration and Fusion" In: Journal of Multimedia, volume 4, 2009, pages: 137-144.
- [2] Larkins, R. Mayo, M., "Adaptive Feature Thresholding for Off-Line Signature Verification", In: Image and vision computing New Zealand, 2008, pages: 1-6.

- [3] Ramachandra, A.C. Pavitra, K.and Yashasvini, K. and Raja, K.B. and Venugopal, K.R. and Patnaik, L.M., "Cross-Validation for Graph Matching based Off-Line Signature Verification", In IDICON 2008, India, 2008, pages: 17-22.
- [4] [4] Kovari, B. Kertesz, Z. and Major, a., "Off-Line Signature Verification Based on Feature Matching: In: Intelligent Engineering Systems, 2007, pages 93-97.
- [5] Porwik P., "The Compact Three Stages Method of the Signatures Recognition", 6 th International Conference on Computer Information Systems and Industrial Management Applications, 2007, pages: 282-287.
- [6] [6] Wei Tian Yizheng Qiao Zhiqiang Ma, "A New Scheme for Off-Line Signature Verification uses DWT and Fuzzy net", In: Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, 2007, pages: 30-35.
- [7] [7] Armand S., Blumenstein, M., Muthukkumarasamy V. "Off-Line Signature and Neural Based Classification", In: Neural Networks, 2006 IJCNN, pages: 684-691.
- [8] [8] Han Ke and Sethi I. K. ,1995. "Handwritten signature retrieval and identification", Pattern Recognition Letter, vol.17,pp.83-90.
- [9] [9] Gongalo Pajares, Jesus, Mahuel de la Cruz, "A wavelet-based image fusion Tutorial", Pattern Recognition Volume 37, Issue 9, September 2004, Elsever Science Inc, pages: 1855-1872.
- [10] [10] Manesh Kokare, P.K. Biswas, and B.N. Chatterji, "Texture Image retrieval using New Rotated Complex Wavelet Filters," IEEE Trans. on systems, man, and Cybernetics-Part B: Cybernetics, vol. 35, no.6, Dec. 2005
- [11] [11] Jean-Luc Starck, Fionn Murtagh, Emmanuel J. Candes, and David L. Donoho, "Gray and color Image Contrast Enhancement by the Curvelet Transform", IEEE Trans. on image processing, vol. 12, no.6, June 2003.

AUTHORS PROFILE

M. S. Shirdhonkar completed his B. E., and M.E. from the Department of Computer Science and Engineering, Shivaji University , Kolhapur, India in the years 1994, 2005 respectively. From 1997-2000, he was worked as lecture in Computer Science Department at JCE, Institute of Technology , Junner, Maharastra, India. In 2000, he joined as a lecturer in the Department of Computer Science at B. L. D. E's. Institute of Engineering and Technology, Bijapur, Karnataka, India, where he is presently holding position of Assistant Professor and doing PhD at S.R.T.M. University, Nanded, Maharastra, India. His research interests include image processing, pattern recognition, and document image retrieval. He is a life member of Indian Society for Technical Education and Institute of Engineers.

Manesh Kokare (S'04) was born in Pune, India, in Aug 1972. He received the Diploma in Industrial Electronics Engineering from Board of Technical Examination, Maharashtra, India, in 1990, and B.E. and M. E. Degree in Electronics from Shri Guru Gobind Singhji Institute of Engineering and Technology Nanded, Maharashtra, India, in 1993 and 1999 respectively, and Ph.D. from the Department of Electronics and Electrical Communication Engineering, Indian Institute of Technology, Kharagpur, India, in 2005. Since June1993 to Oct1995, he worked with Industry. From Oct 1995, he started his carrier in academics as a lecture in the Department of Electronics and Telecommunication Engineering at S. G. G. S. Institute of Engineering and Technology, Nanded, where he is presently holding position of senior lecturer. His research interests include wavelets, image processing, pattern recognition, and Content Based Image Retrieval....

Low Complexity MMSE Based Channel Estimation Technique for LTE OFDMA Systems

Md. Masud Rana¹ and Abbas Z. Kouzani²

¹Department of Electronics and Radio Engineering
Kyung Hee University, South Korea

²School of Engineering, Deakin University, Geelong, Victoria 3217, Australia
Email: mrana928@yahoo.com

Abstract—Long term evolution (LTE) is designed for high speed data rate, higher spectral efficiency, and lower latency as well as high-capacity voice support. LTE uses single carrier-frequency division multiple access (SC-FDMA) scheme for the uplink transmission and orthogonal frequency division multiple access (OFDMA) in downlink. The one of the most important challenges for a terminal implementation are channel estimation (CE) and equalization. In this paper, a minimum mean square error (MMSE) based channel estimator is proposed for an OFDMA systems that can avoid the ill-conditioned least square (LS) problem with lower computational complexity. This channel estimation technique uses knowledge of channel properties to estimate the unknown channel transfer function at non-pilot subcarriers.

Index Terms—Channel estimation, LTE, least-square, OFDMA, SC-FDMA.

I. INTRODUCTION

The 3rd generation partnership project (3GPP) members started a feasibility study on the enhancement of the universal terrestrial radio access (UTRA) in December 2004, to improve the mobile phone standard to cope with future requirements. This project was called evolved-UTRAN or long term evolution [1], [22]. The main purposes of the LTE is substantially improved end-user throughputs, low latency, sector capacity, simplified lower network cost, high radio efficiency, reduced user equipment (UE) complexity, high data rate, and significantly improved user experience with full mobility [2].

3GPP LTE uses orthogonal frequency division multiplexing access (OFDMA) for downlink and single carrier-frequency division multiple access (SC-FDMA) for uplink. SC-FDMA is a promising technique for high data rate transmission that utilizes single carrier modulation and frequency domain equalization. Single carrier transmitter structure leads to keep the peak-to average power ratio (PAPR) as low as possible that will reduced the energy consumption. SC-FDMA has similar throughput performance and essentially the same overall complexity as OFDMA [1]. A highly efficient way to cope with the frequency selectivity of wideband channel is OFDMA. It is an effective technique for combating multipath fading and for high bit rate transmission over mobile wireless channels. In OFDMA system, the entire channel is divided into many narrow subchannels, which are transmitted in parallel, thereby

increasing the symbol duration and reducing the intersymbol-interference (ISI) [2], [4]. Channel estimation (CE) plays an important part in LTE OFDMA systems. It can be employed for the purpose of detecting received signal, improving the capacity of OFDMA systems by cross-layer design, and improving the system performance in terms of symbol error probability (SEP) [4], [5].

A key aspect of the wireless communication system is the estimation of the channel and channel parameters. CE has been successfully used to improve the performance of LTE OFDMA systems. It is crucial for diversity combination, coherent detection, and space-time coding. Improved channel estimation can result: improved signal-to-noise ratio, channel equalization, co-channel interference (CCI) rejection, mobile localization, and improved network performance [1], [2], [3], [18].

Many CE techniques have been proposed to mitigate interchannel interference (ICI) in the downlink direction of LTE systems. In [3], the LS CE has been proposed to minimize the squared differences between the receive signal and estimation signal. The LS algorithm, which is independent of the channel model, is commonly used in equalization and filtering applications. But the radio channel is varying with time and the inversion of the large dimensional square matrix turns out to be ill-conditioned. In [19], Wiener filtering based two-dimensional pilot-symbol aided channel estimation has been proposed. Although it exhibits the best performance among the existing linear algorithms in literature, it requires accurate knowledge of second order channel statistics, which is not always feasible at a mobile receiver. This estimator gives almost the same result as 1D estimators, but it requires higher complexity. To further improve the accuracy of the estimator, Wiener filtering based iterative channel estimation has been investigated [4]. However, this scheme also require high complexity.

In this paper we proposed a channel estimation method in the downlink direction of LTE systems. This proposed method uses knowledge of channel properties to estimate the unknown channel transfer function at non-pilot sub-carriers. These properties are assumed to be known at the receiver for the estimator to perform optimally. The following advantages

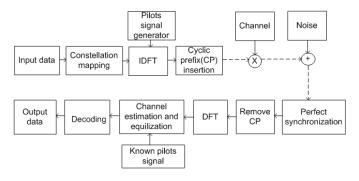


Fig. 1. OFDM transceiver system model.

will be gained by using this proposed method. Firstly, the proposed method avoids ill-conditioned problem in the inversion operation of a large dimensional matrix. Secondly, the proposed method can track the changes of channel parameters, that is, the channel autocorrelation matrix and SNR. However, the conventional LS method cannot track the channel. Once the channel parameters change, the performance of the conventional LS method will degrade due to the parameter mismatch. Finally, the computational complexity of the proposed method is significantly lower than existing LS and Wiener CE method.

We use the following notations throughout this paper: bold face lower and upper case letters are used to represent vectors and matrices, respectively. Superscripts \mathbf{x}^\dagger denote the conjugate transpose of the complex vector \mathbf{x} , diag(x) is the diagonal matrix that its diagonal is vector \mathbf{x} ; and the symbol E(.) denotes expectation.

The remainder of the paper is organized as follows: section II describes LTE OFDMA system model. The proposed channel estimation scheme is presented in section III, and its performance is analyzed in section IV. Section V concludes the work.

II. SYSTEM DESCRIPTION

A. System model

A simplified block diagram of the LTE OFDMA transceiver is shown in Fig.1. At the transmitter side, a baseband modulator transmits the binary input to a multilevel sequences of complex number m(n) in one of several possible modulation formats including binary phase shift keying (BPSK), quandary PSK (QPSK), 8 level PSK (8PSK), 16-QAM, and 64-QAM [1]. CE usually needs some kind of pilot information as a point of reference. CE is often achieved by multiplexing known symbols, so called, pilot symbols into data sequence [15]. These modulated symbols, both pilots and data, are perform a N-point inverse discrete Fourier transform (IDFT) to produce a time domain representation [1]:

$$s(m) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} m(n) e^{\frac{j2\pi nm}{N}},$$
 (1)

where m is the discrete symbols, n is the sample index, and m(n) is the data symbol. The IDFT module output is followed by a cyclic prefix (CP) insertion that completes the digital

stage of the signal flow. A cyclic extension is used to eliminate intersymbol-interference (ISI) and preserve the orthogonality of the tones.

B. Channel model

Channel model is a mathematical representation of the transfer characteristics of the physical medium. These models are formulated by observing the characteristics of the received signal. According to the documents from 3GPP [15], in the mobile environment, a radio wave propagation can be described by multipaths which arise from reflection and scattering. If there are L distinct paths from transmitter to the receiver, the impulse response of the wide-sense stationary uncorrelated scattering (WSSUS) fading channel can be represented as [4]:

$$w(\tau,t) = \sum_{l=0}^{L-1} w_l(t)\delta(\tau - \tau_l), \qquad (2)$$

where fading channel coefficients $w_l(t)$ are the wide sense stationary i.e. $w_l(t) = w(m,l)$, uncorrelated complex Gaussian random paths gains at time instant t with their respective delays τ_l , where w(m,l) is the sample spaced channel response of the lth path during the time m, and $\delta(.)$ is the Dirac delta function. Based on the WSSUS assumption, the fading channel coefficients in different delay taps are statistically independent. Fading channel coefficient is determined by the cyclic equivalent of sinc-fuctions [7]. In time domain fading coefficients are correlated and have Doppler power spectrum density modeled in Jakes [13] and has an autocorrelation function given by [5]:

$$E[w(m,l)w(n,l)^{\dagger}] = \sigma_w^2(l)r_t(m-n) = \sigma_w^2(l)J_0[2\pi f_d T_f(m-n)],$$
 (3)

where w(n,l) is a response of the lth propagation path measured at time n, $\sigma_w^2(l)$ denotes the power of the channel coefficients, f_d is the Doppler frequency in Hertz, T_f is the OFDMA symbol duration in seconds, and $J_0(.)$ is the zero order Bessel function of the first kind. The term f_dT_f represents the normalized Doppler frequency [5].

C. Received signal model

At the receiver, the opposite set of the operation is performed. We assume that the synchronization is perfect. Then, the cyclic prefix samples are discarded and the remaining N samples are processed by the DFT to retrieve the complex constellation symbols transmitted over the orthogonal subchannels. The received signal can be expressed as [5]:

$$r(m) = \sum_{l=0}^{L-1} w(m,l)s(m-l) + z(m), \tag{4}$$

where s(m-l) is the complex symbol drawn from a constellation s of the lth paths at time m-l, and z(m) is the additive white Gaussian noise (AWGN) with zero mean and variance x. After DFT operation, the received signal at pilot

locations is extracted from signal and the corresponding output is represented as follows:

$$R(k) = \sum_{m=0}^{M-1} r(m)e^{\frac{-j2\pi mk}{M}}$$

$$= \sum_{m=0}^{M-1} [w(m,l)s(m-l) + z(m)]e^{\frac{-j2\pi mk}{M}}$$
 (5)

The received signals are demodulated and soft or hard values of the corresponding bits are passed to the decoder. The decoder analyzes the structure of received bit pattern and tries to reconstruct the original signal. In order to achieve good performance the receiver has to know the impact of the channel.

D. OFDMA waveform

The frequencies (sub-carriers) are orthogonal, meaning the peak of one sub-carrier coincides with the null of an adjacent sub-carrier. With the orthogonality, each sub-carrier can be

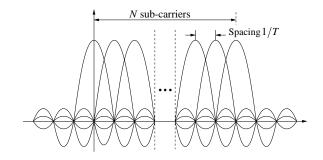


Fig. 2. Orthogonal overlapping spectral shapes for OFDMA system.

demodulated independently without ICI. In OFDM system, the entire channel is divided into many narrow sub-channels, which are transmitted in parallel, thereby increasing the symbol duration and reducing the ISI.

Like OFDM, OFDMA employs multiple closely spaced subcarriers, but the sub-carriers are divided into groups of subcarriers. Each group is named a sub-channel. The sub-carriers that form a sub-channel need not be adjacent. In the downlink, a sub-channel may be intended for different receivers. Finally, OFDMA is a multi-user OFDM (single user) that allows multiple access on the same channel. Despite many benefits of OFDMA for high speed data rate services, they suffer from high envelope fluctuation in the time domain, leading to large PAPR. Because high PAPR is detrimental to user equipment (UE) terminals, SC-FDMA has drawn great attention as an attractive alternative to OFDMA for uplink data transmission.

III. CE PROCEDURE

CE is the process of characterizing the effect of the physical medium on the input sequence. The aim of most CE algorithm is to minimize the mean squared error (MSE), while utilizing as little computational resources as possible in the estimation process [2], [4]. CE algorithms allow the receiver to approximate the impulse response of the channel

and explain the behavior of the channel. This knowledge of the channel's behavior is well-utilized in modern mobile radio communications. One of the most important benefits of channel estimation is that it allows the implementation of coherent demodulation. Coherent demodulation requires the knowledge the phase of the signal. This can be accomplished by using channel estimation techniques. Once a model has

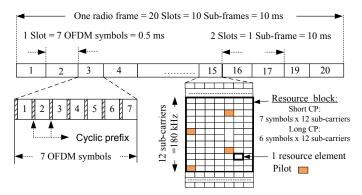


Fig. 3. OFDMA generic frame structure.

been established, its parameters need to be estimated in order to minimize the error as the channel changes. If the receiver has a priori knowledge of the information being sent over the channel, it can utilize this knowledge to obtain an accurate estimate of the impulse response of the channel.

In LTE, like many OFDMA systems, known symbols called training sequence, are inserted at specific locations in the time frequency grid in order to facilitate channel estimation [10], [15]. As shown in Fig. 3, each slot in LTE downlink has a pilot symbol in its seventh symbol [6] and LTE radio frames are 10 msec long. They are divided into 10 subframes, each subframe 1 msec long. Each subframe is further divided into two slots, each of 0.5 msec duration. The subcarrier spacing in the frequency domain is 15 kHz. Twelve of these subcarriers together (per slot) is called a physical resource block (PRB) therefore one resource block is 180 kHz [2], [3], [6]. Six resource blocks fit in a carrier of 1.4 MHz and 100 resource blocks fit in a carrier of 20 MHz. Slots consist of either 6 or 7 ODFM symbols, depending on whether the normal or extended cyclic prefix is employed [10], [15], [17].

Channel estimates are often achieved by multiplexing training sequence into the data sequence [18]. These training symbols allow the receiver to extract channel attenuations and phase rotation estimates for each received symbol, facilitating the compensation of channel fading envelope and phase. General channel estimation procedure for LTE OFDMA system is shown in Fig. 4. The signal $\bf S$ is transmitted via a time-varying channel $\bf w$, and corrupted by an additive white Gaussian noise (AWGN) $\bf z$ before being detected in a receiver. The reference signal $\bf w_{est}$ is estimated using LS , Wiener based, or proposed method. In the channel estimator, transmitted signal $\bf S$ is convolved with an estimate of the channel $\bf w_{est}$. The error between the received signal and its estimate is

$$\mathbf{e} = (\mathbf{r} - \mathbf{r}_1). \tag{6}$$

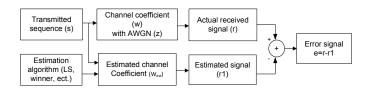


Fig. 4. General channel estimation procedure.

The aim of most channel estimation algorithms is to minimize the mean squared error (MMSE), while utilizing as little computational resources as possible in the estimation process. The equation (4) can be written as vector notation as [1]:

$$\mathbf{r} = \mathbf{S}\mathbf{w} + \mathbf{z},\tag{7}$$

where $\mathbf{r} = (r_0, r_1,, r_{L-1})^{\dagger}$, $\mathbf{S} = diag(s_0, s_1,, s_{L-1})$, $\mathbf{w} = (w_0, w_1,, w_{L-1})^{\dagger}$, and $\mathbf{z} = (z_0, z_1,, z_{L-1})^{\dagger}$. The least-square estimate of such a system is obtained by minimizing square distance between the received signal and its estimate as [3]:

$$\mathbf{J} = (\mathbf{Sr} - \mathbf{w})^2 = (\mathbf{r} - \mathbf{Sw})(\mathbf{r} - \mathbf{Sw})^{\dagger}. \tag{8}$$

We differentiate this with respect to \mathbf{w}^{\dagger} and set the results equal to zero to produce [3]:

$$\mathbf{w}_{LS} = (\alpha \mathbf{I} + \mathbf{S} \mathbf{S}^{\dagger})^{-1} \mathbf{S}^{\dagger} \mathbf{r}, \tag{9}$$

where α is regularization parameter and has to be chosen such that the resulting eigenvalues are all defined and the matrix $(\alpha \mathbf{I} + \mathbf{SS}^{\dagger})^{-1}$ is the least perturbed. Where the channel is considered as a deterministic parameter and no knowledge on its statistics and on the noise is needed. The LS estimator is computationally simple but problem that is encountered in the straight application of the LS estimator is that the inversion of the square matrix turns out to be ill-conditioned. So, we need to regularize the eigenvalues of the matrix to be inverted by adding a small constant term to the diagonal [3]. If the transmitted signal is more random, the performance of the LS method is significantly decrease. Also the LS estimate of \mathbf{w}_{est} is susceptible to Gaussian noise and inter-carrier interference (ICI). Because the channel responses of data subcarriers are obtained by interpolating the channel responses of pilot subcarriers, the performance of OFDM system based on combtype pilot arrangement is highly dependent on the rigorousness of estimate of pilot signals. The successful implementation of the LS estimator depends on the existence of the inverse matrix $(SS^{\dagger})^{-1}$. If the matrix (SS^{\dagger}) is singular (or close to singular), then the LS solution does not exist (or is not reliable).

To improve the accuracy of the estimator, Wiener filtering based iterative channel estimation has been investigated [4], [7]:

$$\mathbf{w}_{est} = \mathbf{R}_{ww} \mathbf{F}^{\dagger} \mathbf{S}^{\dagger} [(\mathbf{SFR}_{ww} \mathbf{F}^{\dagger} \mathbf{S}^{\dagger}) + x \mathbf{I}]^{-1} \mathbf{w}_{ls}$$
(10)

where \mathbf{R}_{ww} is the autocovariance matrix of \mathbf{w} , \mathbf{F} is the DFT matrix, and x denotes the noise variance. However, this scheme also requires higher complexity.

IV. PROPOSED MMSE BASED CE TECHNIQUE

The equation (7) can we rewritten as [22]:

$$\mathbf{w}_1 = \frac{\mathbf{r}}{\mathbf{S}} + \frac{\mathbf{z}}{\mathbf{S}}$$
$$= \mathbf{w}_2 + \mathbf{z}_1, \tag{11}$$

where actual channel value is $\mathbf{w}_2 = \mathbf{r}/\mathbf{S}$, noise contribution $\mathbf{z}_1 = \mathbf{z}/\mathbf{S}$, and \mathbf{w}_1 is the result of direct estimated channel. The proposed channel estimation is

$$\mathbf{w}_{prop} = \sum_{k=0}^{L-1} \mathbf{a}_k^{\dagger} \mathbf{w}_1(k)$$

$$= \sum_{k=0}^{L-1} \mathbf{a}_k^{\dagger} [\mathbf{w}_2(k) + \mathbf{z}_1(k)]$$

$$\mathbf{w}_{prop} = \mathbf{a}^{\dagger} \cdot \mathbf{w}_3, \tag{12}$$

where $\mathbf{a}_k = (a_0, a_1..., a_{L-1})^\dagger$ is the column vector filter coefficients, and $\mathbf{w}_3 = \sum_{k=0}^{L-1} [\mathbf{w}_2(k) + \mathbf{z}_1(k)]$. The mean square error (MSE) for the proposed LTE channel estimation is $\mathbf{J} = (\mathbf{w} - \mathbf{w}_{prop})^2$. In order to calculate the optimal coefficient, taking the expectation of MSE and partial derivative with respect to channel coefficient:

$$\frac{\partial E(\mathbf{J})}{\partial \mathbf{a}^{\dagger}} = \frac{\partial}{\partial \mathbf{a}^{\dagger}} (E[(\mathbf{w} - \mathbf{w}_{prop})(\mathbf{w} - \mathbf{w}_{prop})^{\dagger}]). \tag{13}$$

Now putting the value of $\mathbf{w}_{prop} = \mathbf{a}^{\dagger} \mathbf{w}_3$ into the above equation to produce:

$$\frac{\partial E(\mathbf{J})}{\partial \mathbf{a}^{\dagger}} = \frac{\partial}{\partial \mathbf{a}^{\dagger}} (E[(\mathbf{w} - \mathbf{a}^{\dagger} \mathbf{w}_{3})(\mathbf{w} - \mathbf{a}^{\dagger} \mathbf{w}_{3})^{\dagger}])$$

$$= \frac{\partial}{\partial \mathbf{a}^{\dagger}} (E[(\mathbf{w} - \mathbf{a}^{\dagger} \mathbf{w}_{3})(\mathbf{w}^{\dagger} - \mathbf{a} \mathbf{w}_{3}^{\dagger})])$$

$$= \frac{\partial}{\partial \mathbf{a}^{\dagger}} (E[\mathbf{w} \mathbf{w}^{\dagger} - \mathbf{a}^{\dagger} \mathbf{w}_{3} \mathbf{w}^{\dagger} - \mathbf{a} \mathbf{w}_{3}^{\dagger} \mathbf{w} + \mathbf{a}^{\dagger} \mathbf{w}_{3} \mathbf{w}_{3}^{\dagger} \mathbf{a}])$$

$$= E[-\mathbf{w}_{3} \mathbf{w}^{\dagger} + \mathbf{w}_{3} \mathbf{w}_{3}^{\dagger} \mathbf{a}]. \tag{14}$$

Now putting the partial derivative equal to zero in the above equation and after some manipulations we get the coefficient as:

$$\mathbf{a} = E[(\mathbf{w}_{3}\mathbf{w}^{\dagger})](E[(\mathbf{w}_{3}\mathbf{w}_{3}^{\dagger})])^{-1}$$

$$= [E(\mathbf{w}_{2} + \mathbf{z}_{1})\mathbf{w}^{\dagger}][E((\mathbf{w}_{2} + \mathbf{z}_{1})(\mathbf{w}_{2} + \mathbf{z}_{1})^{\dagger})]^{-1}$$

$$= [E(\mathbf{w}_{2}\mathbf{w}^{\dagger} + \mathbf{z}_{1}\mathbf{w}^{\dagger})][E((\mathbf{w}_{2} + \mathbf{z}_{1})(\mathbf{w}_{2}^{\dagger} + \mathbf{z}_{1}^{\dagger}))]^{-1}$$

$$= E(\mathbf{w}_{2}\mathbf{w}^{\dagger} + \mathbf{z}_{1}\mathbf{w}^{\dagger})[E(\mathbf{w}_{2}\mathbf{w}_{2}^{\dagger} + \mathbf{z}_{1}\mathbf{w}_{2}^{\dagger} + \mathbf{w}_{2}\mathbf{z}_{1}^{\dagger} + \mathbf{z}_{1}\mathbf{z}_{1}^{\dagger})]^{-1}.$$
(15)

In this paper we assume that mean of the AWGN is zero i.e. $E(\mathbf{z}) = 0$ and variance is x i.e. $E(\mathbf{z}\mathbf{z}^{\dagger}) = x$. So, the above equation is simplified as:

$$\mathbf{a} = E(\mathbf{w}_{2}\mathbf{w}^{\dagger})[E(\mathbf{w}_{2}\mathbf{w}_{2}^{\dagger}) + E(\mathbf{z}_{1}\mathbf{z}_{1}^{\dagger})]^{-1}$$

$$= E(\mathbf{w}_{2}\mathbf{w}^{\dagger})[E(\mathbf{w}_{2}\mathbf{w}_{2}^{\dagger}) + x]^{-1}$$

$$= \mathbf{w}_{cross} * (\mathbf{W}_{auto} + x)^{-1}, \qquad (16)$$

where $\mathbf{w}_{cross} = E(\mathbf{w}_2\mathbf{w}^{\dagger})$ and $\mathbf{W}_{auto} = E(\mathbf{w}_2\mathbf{w}_2^{\dagger})$ are the channel cross-correlation vector and autocorrelation matrix respectively. Now putting this filter coefficient value in equation (12), we get the final channel estimation formula as:

$$\mathbf{w}_{prop} = [\mathbf{w}_{cross} * (\mathbf{W}_{auto} + x)^{-1}] \mathbf{w}_3. \tag{17}$$

V. COMPLEXITY COMPARISON

The complexity of CE is of crucial importance especially for time varying wireless channels, where it has to be performed periodically or even continuously. For this proposed estimator, the main contribution to the complexity comes from the term $[\mathbf{w}_{cross}*(\mathbf{w}_{auto}+x)^{-1}]$. The variance of the AWGN is precalculated and added with the autocorrelation matrix. Thus, only one run-time matrix inversion is required. Also \mathbf{w}_3 is pre-calculated column vector. Table I summarizes the computational complexity of the proposed and existing channel estimation methods. It shows that the proposed CE algorithm has lower complexity than existing methods.

TABLE I COMPUTATIONAL COMPLEXITY OF ALGORITHMS

Operation	LS method	Wiener method	Proposed method
Matrix inversion	1	1	1
Multiplication	3	6	2
Addition	1	1	1

VI. CONCLUSION

In this paper, we present a MMSE based channel estimation method for LTE OFDMA systems and compared the performance with the LS and Wiener based filtering method. This proposed channel estimation method uses knowledge of channel properties to estimate the unknown channel transfer function at non-pilot sub-carriers. It can well solve the ill-conditioned least square (LS) problem and track the changes of channel parameters with low complexity .

ACKNOWLEDGMENT

The author would like to thanks Prof. Dr. Jinsang Kim.

REFERENCES

- B. Karakaya, H. Arslan, and H. A. Cirpan, "Channel estimation for LTE uplink in high doppler spread," *Proc. WCNC*, pp. 1126-1130, April 2008.
- [2] J. Berkmann, C. Carbonelli, F.Dietrich, C. Drewes, and W. Xu, "On 3G LTE terminal implementation standard, algorithms, complexities and challenges," *Proc. Int. Con. on Wireless Communications and Mobile Computing*, pp. 970-975, August 2008.
- [3] A. Ancora, C. Bona, and D. T. M. Slock, "Down-sampled impulse response least-squares channel estimation for LTE OFDMA," Proc. Int. Con. on Acoustics, Speech and Signal Processing, Vol. 3, pp. 293-296, April 2007
- [4] L. A. M. R. D. Temino, C. N. I Manchon, C. Rom, T. B. Sorensen, and P. Mogensen, "Iterative channel estimation with robust wiener filtering in LTE downlink," *Proc. Int. Con. on Vehicular Technology Conference*, pp. 1-5, September 2008.
- [5] S. Y. Park, Y.Gu. Kim, and C. Gu. Kang, "Iterative receiver for joint detection and channel estimation in OFDM systems under mobile radio channels," *Vehicular Technology, IEEE Transactions*, Vol. 53, Issue 2, pp. 450-460, March 2004.

- [6] J. Zyren, "Overview of the 3GPP long term evolution physical layer," Dr. Wes McCoy, Technical Editor, 2007.
- [7] J. J. V. D. Beek, O. E. M. Sandell, S. K. Wilsony, and P. O. Baorjesson, "On channel estimation in OFDM systems," *Proc. Int. Con. on Vehicular Technology Conference*, vol. 2, pp. 815-819, July 1995.
- [8] D. G. Manolakis, D.Manolakis, V. K. Ingle, and S. M. Kogon, "Statistical and adaptive signal processing," *McGraw-Hill*, 2000.
- [9] L. J. Cimini, and Jr, "Analysis and simulation of a digital mobile channel using orthogonal frequency division multiplexing," *IEEE Transactions Communication*, vol. 33, no. 7, pp. 665-675, July 1985.
- [10] "Requirements for EUTRA and EUTRAN," 3GPP TR 25.913 V7.3.0, 2006.
- [11] H. G. Myung, J. Lim, and D. J. Goodman, "Single carrier FDMA for uplink wireless transmission," *IEEE Vehicular Technology Magazine*, vol. 1, no. 3, pp. 30-38, September 2006.
- [12] K. Han, S. Lee, J. Lim, and K. Sung, "Channel estimation for OFDM with fast fading channels by modified Kalman filter," Consumer Electronics, IEEE Transactions on, vol. 50, no. 2, pp. 443-449, May 2004.
- [13] W. Jakes, and D. Cox, "Microwave mobile communications," Wiley-IEEE Press, 1994.
- [14] O. Edfors, M. Sandell, J. V. D. Beek, and S. Wilson, "OFDM channel estimation by singular value decomposition," *IEEE Transactions on Communications*, vol. 46, no. 7, pp. 931-939, July 1998.
- [15] "Technical specification group radio access networks; deployment aspects," 3rd Generation Partnership Project, Tech. Rep. TR 25.943, V7.0.0, June 2007
- [16] H.G. Myung, J. Lim, and D. J. Goodman, "Peak to average power ratio for single carrier FDMA signals," *Proc. PIMRC*, 2006.
- [17] S. Maruyama, S. Ogawa, and K.Chiba, "Mobile terminals toward LTE and requirements on device technologies," *Proc. Int. Con. on VLSI Circuits, IEEE Symposium on*, pp. 2-5, June 2007.
- [18] M.H. Hsieh, and C.H. Wei, "Channel estimation for OFDM systems based on comb-type pilot arrangement in frequency selective fading channels," *Consumer Electronics, IEEE Transactions on*, vol. 44, issue 1, pp. 217-225, Feb. 1998.
- [19] P. Hoeher, S. Kaiser, and P. Robertson, "Two-dimensional pilot-symbolaided channel estimation by wiener filtering," *Proc. Int. Con. on Acoustics, Speech, and Signal Processing*, pp. 1845-1848, vol.3, April 1997
- [20] S. H. Han, and J. H. Lee, "An overview of peak-to-average power ratio reduction techniques for multicarrier transmission," *IEEE Wireless Communications*, April 2005.
- [21] H. H. Roni, S. W. Wei, Y. H. Jan, T. C. Chen, and J.H. Wen, "Low complexity qSIC equalizer for OFDM System," Proc. Int. Con. on Symposium on Consumer Electronics, pp. 434-438, May 2005.
- [22] M. M. Rana, J. Kim, and W. K. Chow, "Low complexity downlink channel estimation for LTE systems," *Proc. Int. Con. on Advanced Communication Technology*, pp. 1198-1202, Febuary 2010.

SURVEY: RTCP FEEDBACK IN A LARGE STREAMING SESSIONS

Adel Nadhem Naeem¹, Ali Abdulqader Bin Salem² Mohammed Faiz Aboalmaaly³ and Sureswaran Ramadass⁴
National Advanced IPv6 Centre
Universiti Sains Malaysia
Pinang, Malaysia

Abstract—RTCP has limitation with scalability for large streaming sessions; because of the limitation of the bandwidth space that given to RTCP reports. Many researchers studied and still studying how to solve this limitation, and most of the researchers come out with tree structure as a solution but in a different ways.

Keywords- RTCP/RTP; Scalability; Large Streaming Sessions

I. INTRODUCTION

Communication is everywhere in our life and as a part of it multimedia communication like, VoIP, multimedia conferencing, teleconferencing, video surveillance, satellite communication, etc. Multimedia communication is mainly using Real time protocol (RTP) that works together with Real time control protocol (RTCP) to do transfer data that depend on real time like video or audio over networks. RTCP is used to monitor RTP-Packets and reports feedback [1]. The main function of RTCP is to transmit periodically the sender and receiver reports to all members in RTP/RTCP. These reports allow a host to know if a problem exists or not and if the problem is local or global [2]. RTCP like other protocols and techniques facing a lot of problems, one of the important problems is RTCP scalability. [3] Increasing the number of hosts in RTP/RTCP sessions caused some problems under the name of the scalability problems; the problems can be the feedback delay, storage problem, flood of initial/bye RTCP reports, etc.

II. SCALABILITY OF RTCP

RTP is a real time transmission protocol of audio and video, which provide several functions that help the transmission of audio and video such as [4]. Identification of payload data type, give a Sequence numbering to detect packet loss and to order packets, Time stamping so that data is played out at the right speeds [5]. RTCP is the attached protocol to RTP, and its working by sending/receiving reports, and it has several kinds of reports the main reports in RTCP are the sender report (SR) and receiver report (RR). Both include performance statistics on the total number of packets loss since the beginning of transmission, the fraction of packet loss in the interval between sending this feedback report and sending the previous one, the highest sequence number received, jitter, and other delay measurements to calculate the round-trip feedback delay time.

The SR provides more statistics summarizing data transmission from the sender, e.g. timestamps, count of RTP data packets, and number of payload octet's transmitted [4].

- RRs are used mainly in sender-based adaptive applications (The packet loss parameter in the RRs has been used as an indicator of congestion in the network).
- The SR is useful in lip-synchronization (inter-media synchronization) and in calculating transmission bit rates.

What does scalability mean when it uses with RTCP term, increase the number of hosts in one session then increase the number of RTCP report, which means RTCP scalability. RTCP is kind of reports' protocol that sends/receive reports from/to hosts in one session. These reports limited to bandwidth size, RTCP given 5% from the whole session bandwidth size. RTCP has two types of the report: Sender report and Receiver report. Sender report uses 75% of RTCP bandwidth size while Receiver report uses 25%. The limitation of bandwidth size of RTCP makes it control the interval of sending receiving reports, increasing the number of hosts caused to increase the interval of sending receiving time and that makes the reports useless [6].

III. CHALLENGES OF RTCP SCALABILITY

Many researchers studied and wrote about RTCP problems, and the most important problems are about the scalability of RTCP and reports' feedback in a large streaming session. The scalability in RTCP faces many problems when it comes to a group of thousands of users. Some of these problems are addressed in [2 ... 11]. The first serious study done by El-Merakby, with three main researches 1998 "A Scalability Scheme for the Real-time Control Protocol", 2000 "Design and Performance of a Scalable Real Time Control Protocol: Simulations and Evaluations", and 2005 "Scalability improvement of the real time control protocol" where she studied RTCP problems with scalability and suggested a solution by divide the large session group to small session groups [4, 8, 9].

A. Feedback delay challenge

One of the important challenges is the feedback delay and caused by increasing the group size, because of the limitation of the bandwidth size the RTCP reporting interval increased which decreases the significance and value of the feedback, and then the feedback reports either send rarely or not at all [3, 4, 8, 9].

B. Storage challenge

The group size could be known if every member stores a count of distinct for every member it heard during the session using the unique Synchronization Source identifier (SSRC) found in the RTP header [3, 4, 8, 9].

C. Multicasting RRs to the whole group (bandwidth effect)

Every member in the session group will multicast RRs (Receiver Report) to all other which are not senders and that causes a load at every member processing and Congestion will happen because of members increase then RR increase also[4, 8, 9, 10].

D. Initial/bye flood challenge

If many members join/leave the session at the same time, a flood of join/Bye packets will happen and congestion in the network may occur, especially at members who have low bandwidth links [3, 4, 5, 8, 9].

El-Merakby tried to explain that the normal case for RTCP feedback reports are multicast mainly for receivers to calculate the group size and thus compute their RTCP reporting interval, and the suggested solution is saying that the members do not need to compute the whole size of the multicast group and RRs are not multicast, and to divide the big session to many groups. The proposed structure is called S-RTCP, and shown in Fig 1, explains how members organize dynamically in a multi-level hierarchy of local regions.

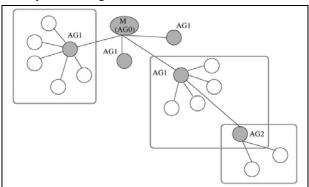


Figure 1. Structure of El-Merakby scheme [9].

Each region has an aggregator (AG). Each member sends the RR feedback to its AG which gathers and aggregates statistics from these reports which is passed to a manager or to another AG level. Additional statistics are computing by the manager to evaluate the transmission quality and to estimate the regions which suffer from congestion. Time-to-Live (TTL) field in the IP header is used by the scheme to build the multilevel hierarchy with locally scoped regions [4, 8, 9].

The following are the advantages of using the scheme in large RTCP groups [4, 8, 9]:

- Resolving the storage scalability problem: Members do not need to store the state of the distinct member in the group because they are in a different small group size.
- Timely reporting of feedback reports: Feedback reports become more useful because the number of members became less.
- Effective use of the bandwidth: the formation of local regions where RRs are not multicast but are sent with limited scope and not global scope decrease the number of RRs.
- Decrease in the number of redundant reports: the total number of redundant RRs, which used to be multicast, is decreased, because the measurements in RRs are aggregated into AGRs summarizing the quality of the received data.

In the other side, another researcher was interesting in the same area, Julian from University of Cambridge. He published an article with title "An Extensible RTCP Control Framework for Large Multimedia Distributions" in 2003. Julian thought that is two serious challenges with RTCP and they are: the growing of using unidirectional and asymmetric broadcast architectures, and the second challenge: per-receiver RTCP reporting frequency diminishes prohibitively due to the bandwidth-sharing algorithm [2].

A. The growing deployment of unidirectional and asymmetric broadcast architectures challenge

In RTP/RTCP, the data and control share a many-to-many communication channel, such as that provided by IP multicast [11]. The unidirectional and asymmetric broadcast architectures have problems with these issues; instead the channel allows not only the bidirectional flow of communication from sources to receivers and vice-versa, but also direct receiver-to-receiver communication over a single channel [2, 11].

B. Per-receiver RTCP reporting frequency diminishes prohibitively due to the bandwidth-sharing algorithm

RTCP is keeping the frequency of reports inversely proportional to the number of members. And because of that RTCP institutes a bandwidth-sharing algorithm that divides the resources of the control channel among members' group. The standard bandwidth-sharing algorithm used by RTCP expects that as groups grow in size, the frequency of individual feedback reports will decrease [2]. This problem is the same challenge that introduced by El-Merakby, 1998 with a challenge title Feedback delay challenge [4].

To solve these challenges, two new schemes are devised that are complementary to the existing RTCP feedback algorithm and influence the unique characteristics of summaries to efficiently scale the feedback of the unicast backchannel for large groups. The two schemes that influence summarization to scale the backchannel: biasing and hierarchical aggregation [2].

- The technique of biasing provides preferential treatment to the feedback of one or more groups of receivers.
- The technique of hierarchical aggregation supports the existence of multiple summarization nodes throughout a collection topology, which distributes the load and bandwidth usage of summarization, and in turn lends much-needed support to heterogeneous topologies as well as to frequency-driven applications.

In 2005, another researcher was interesting in scalability of RTCP, but instead of studying the main RTCP protocol. He decided to study S-RTCP (Scalable Real Time Protocol) that invented by El-Marakby [4, 8,9], Elramly published two papers "Scalability Solutions For Multimedia Real-Time Control Protocol (PDPTA'06)" 2005, and "Analysis, Design, and Performance Evaluation of MS-RTCP: More Scalable Scheme for the Real-Time Control Protocol" 2005. Elramly introduced a new protocol MS-RTCP (More Scalable Real Time Control Protocol). MS-RTCP scheme is based on a hierarchical structure, distributed management, and EL-Marakby scheme. The idea of El-Marakby scheme is depending on a tree-based hierarchy of report summarizers. The tree leafs (nodes) in the session send RRs to some node that acting as AG (aggregator), collects and summarizes these reports. The summaries result then passed up to the next highest level in the tree, until finally they reach the sender or some other appropriate feedback point. The summarization scheme is most useful when the nodes at one level of a sub-tree see similar network performance. El-Marakby uses network hop counts to a summarizer, measured through Time To Live (TTL), to group hosts together in the tree. She also proposed a dynamic scheme for building the tree [4, 8, 9, 12, 13]. The most important introduced problems by Elramly to El-marakby scheme (S-RTCP):

- Fault tolerance is not guaranteed: When any AG is crashed or has left the RTP session, all the children in its region search for other AG. This will affect the convergence time during this interval, and will make an old child to a new one.
- Load balancing is not guaranteed: The load balancing depends on the maximum number of children with which the AGs deals. This is not sufficient, because if we suppose that the maximum number of each AG is 100 children, we may find AG has 90 children and another has 10 children.
- The structure of the model depends on the central processing unit (manager). Hence, if the manager is crashed there is no other unit that can take place. The model in this case will become unstable (failed in worst case).
- 4. Overkill the small groups, which the IP telephone is mainly dealing with. This is due to the condition of low children number per AG that was put after the model evaluation [8].
- 5. Election of LAN Aggregator (LAG) is not sufficient. Source Description (SDES) items [1] about any multicast group member will take a long time to access it.

The MS-RTCP managers join the control multicast group, while the MS-RTCP children join the data multicast group. Each group of children constructs a region which is controlled by a manager. Each child should know its region before sharing the RTP session. The Load Balancing Manager (LBM) accomplishes this target by testing the real position of each child and the real number of children per region. The scheme structure is shown in Fig. 2.

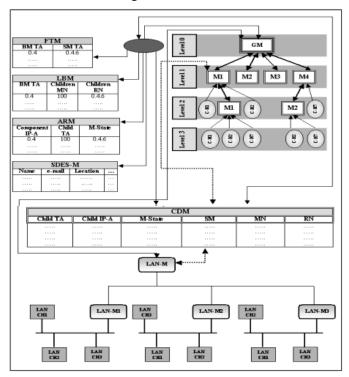


Figure 2. General view of Elramly schem [13].

The suggested solution and the proposed protocol MS-RTCP:

- 1. Fault tolerance guarantee: any control component in the MS-RTCP has a spare one. If one or more scheme components fail, can replace the failed one with its spare one by the decision from the GM (or FTM) until the failed one is fixed and the normal situation is re-established.
- 2. Load balancing guarantee: the decision for receiving a new child in the RTP session depends on the real number of children per scheme manager. Consequently, if any new child joins a session, it is told with the best manager taking in consideration the manager load (real children number) and the new child position.
- 3. The basic idea of the MS-RTCP is based on the management distribution. So, the central management processing is eliminated, as the scheme has one manager for each management process.
- 4. The maximum number of children per manager reaches the upper limit at which the RTP session is working safely (some hundreds or more). Hence, if a small group joins the RTP session, the MS-RTCP will be transformed to the simple RTCP view with one manager and it's spare. The other MS-RTCP managers will be found, but with minimal overhead (neglected values).

- The LAN Manager (LAN-M) has its pre-determined spare component. So, in this scheme, no need to elect another LAN-M when the basic one fails.
- 6. The GM can access any data about any MS-RTCP entity by requesting the SDESM (or by CDM in case of any problem happened for the SDES-M).

In Brno University of Technology, Czech Republic, 2007, new researchers work with scalability of RTCP. Dan Komosny, and Vit Novotny, started to study scalability of RTCP and its challenges. They published three papers together, "Tree Structure for Source-Specific Multicast with Feedback Aggregation" 2007, "Optimization of Large-Scale RTCP Feedback Reporting in Fixed and Mobile Networks" 2007, and "Large-Scale RTCP Feedback Optimization" 2008. The researchers find that RTCP in a large session causes delays in sending feedback data from each receiver, and to solve this problem they proposed a hierarchical structure and new protocol called Tree Transmission Protocol (TTP). Their work more to be extended to Julian 2003[2], where they solve some of the hierarchical structure that proposed by Julian 2003. The Fig. 3 is shown the tree structure [14, 15, 16, 17].

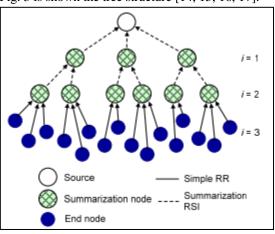


Figure 3. Tree topology of the RTCP feedback network [17].

To solve the problem of hierarchical structure organization new protocol has been proposed TTP (Tree Transmission Protocol), this protocol is quite a flexible and robust protocol use to organizing the hierarchical tree overlays. It can be used for any hierarchically organized protocols. It can work for simple hierarchical tree overlays as well as for large-scale overlays with many hierarchical levels. The fig 4 shows TTP position [15, 16].

Media transmission	Feedback transmission		Feedback structure management			
RTP	RTCP		TTP			
UDP		TCP				
IP						

Figure 4. Position of TTP to related protocols [16].

Dan Komosny, and Vit Novotny come out with:

- 1. Solve round trip delay by suggestion tree structure.
- The problem of the RTCP feedback tree establishment was solved.
- 3. The method for finding the nearest summarization node in the IP network structure was designed.
- 4. To manage the tree the new protocol (TTP) was designed and specified.

Shaahin Shahbazi comes with two papers about RTCP scalability, "A new design for improvement of scalable-RTCP" 2009, and "Error Resistant Real-Time Transport Control Protocol" 2009. As Elramly 2005, he preferred to deal with El-Marakby protocol (Scalable Real Time Control protocol) SRTCP. Shahbazi explained about the challenges associated with S-RTCP, and proposed different approaches to solve the challenges [18, 19, 4, 8, 9].

A. The S-RTCP challenges

- Congestion: If the number of AGs becomes very big of AG-0, congestion may occur at the links connected to AG-0.
- Overload: because of the same reason overload may happen.
- Lack of error-tolerance: S-RTCP design is quite vulnerable to any sort of malfunctions within AG-0.

B. Proposed Design for Stability Improvement

problems are caused due to the singularity of AG-0 in S-RTCP's design and also the fact that no precautions have been taken into account in case AG-0 fails [18]. See Fig. 5

- The number of AG-0 nodes: fix the amount of AG-0s
- Mirrored tasks versus split tasks: all AG-0s receive the same information.
- Existence of pre-assigned AG-0: new design will allow two or more AG-0s to operate within the session, in order to provide stability.

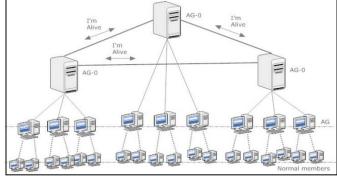


Figure 5. The architecture of the modified version of S-RTCP [18]

Shahbazi was designed a new proposed scheme ER-RTCP (Error Resistant Real-Time Transport Control Protocol). Modifications included designing the multi-manager scheme,

improving parent-seeking procedures, reducing distribution of request packets, reforming the design to be independent of TTL, adding methods to check on sanity of manager nodes. This study considered packet loss ratio of below 2% as desirable [18, 19]. See Fig. 6.

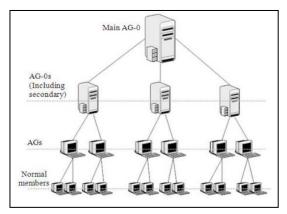


Figure 6. The architecture of ER-RTCP [19]

IV. COMPARISON BETWEEN RESEARCHERS' WORK

The comparison can be show in Table 1.

V. SUMMARY

RTCP protocol has some problems and many researchers studied to solve it specially the scalability issue, the main problem was with the limitation of bandwidth space that given to RTCP reports, which all the researcher come with tree structure as a solution for it but in different ways.

TABLE I. COMPARISON BETWEEN RESEARCHERS' WORK

Researcher	Period of studies	Problems that solved	New suggested protocol
El-Marakby	1998, 2000, 2003, 2005	 The research studied RTCP. Feedback delay. Increasing storage state. Multicasting RRs congestion. Flood (Initial/Bye). Come out with tree structure. 	S-RTCP
Julian Chesterfield	2003	 The research studied RTCP. Unidirectional and asymmetric broadcast architectures problem. Bandwidth-sharing algorithm. Come out with tree structure. 	-
El-Ramly	2005	1- The research studied S-RTCP. 2- Fault tolerance is not guaranteed. 3- Load balancing is not guaranteed. 4- The structure of the model depends on the central processing unit. 5- Overkill the small groups. 6- Election of LAN Aggregator (LAG) is not sufficient. 7- Source Description (SDES) items. 8- Come out with tree structure.	MS-RTCP
Dan Komosny, & Vit Novotny	2007, 2008	 The research studied RTCP. Round trip delay. The RTCP feedback tree establishment. Finding the nearest summarization node in the IP network structure. Come out with tree structure. 	ТТР
Shaahin	2009	1- The research studied S-RTCP. 2- Congestion. 3- Overload. 4- Lack of error-tolerance. 5- Come out with tree structure.	ER-RTCP

ACKNOWLEDGMENT

I would like to thank National Advanced IPv6 Center (NAv6), Universiti Sains Malaysia for their support that enabled me to complete this work.

REFERENCES

- H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, "RTP A Transport Protocol for Real-time Applications," RFC 3550 (STD 64), July 2003.
- [2] J. Chesterfield, E. M. Schooler "An extensible RTCP control framework for large multimedia distributions", 2003.
- [3] J. Rosenberg, H. Schulzrinne, Timer reconsideration for enhanced RTP scalability, Proceedings of IEEE Infocom'98, San Francisco, CA, 1998, pp. 233–241.
- [4] R. El-Marakby, D. Hutchison, A scalability scheme for the real-time control protocol, Proceedings of IFIP TC-6 Eighth International Conference on High Performance Networking (HPN'98), Vienna, 1998, pp.153–168.
- [5] J. Rosenberg, H. Schulzrinne, Timer reconsideration for enhanced RTP scalability, Proceedings of IEEE Infocom'98, San Francisco, CA, 1998, pp. 233–241.
- [6] T. Friedman, R. Caceres, A. Clark "RTCP Reporting Extensions", RFC 3611, 2003.
- [7] J. Rosenberg, H. Schulzrinne, New Results in RTP Scalability, IETE, 1997, Internet draft: draft-ietf-avt-byerecon-00.ps
- [8] R. El-Marakby, Design and performance of a scalable real time control protocol: simulations and evaluations, Proceedings of the Fifth IEEE Symposium on Computers and Communications (ISCC'2000), Antibes-Juan Les Pins, France, 2000, pp.119–124.
- [9] Randa El-Marakby, David Hutchison: Scalability improvement of the real-time control protocol, Computer Communications 28 (2005) 136– 149
- [10] B. Aboba, Alternatives of enhancing RTP scalability, IETF, 1996, Internet draft: draftaboba-rtpscale-02.txt.
- [11] S. Deering. Host extensions for IP multicasting. Request for Comments 1054, Internet Engineering Task Force, Aug. 1989
- [12] O. Essa, N. El-Ramly, and H. Harb, "SCALABILITY SOLUTIONS FOR MULTIMEDIA REAL-TIME CONTROL PROTOCOL (PDPTA'06), 2005.
- [13] N. Elramly, A. Habib, O. Essa, and H. Harb, "Analysis, Design, and Performance Evaluation of MS-RTCP: More Scalable Scheme for the Real-Time Control Protocol," Journal of Universal Computer Science, vol. 11, pp. 874-897, 2005.
- [14] V. Novotný, D. Komosný, "Optimization of Large-Scale RTCP Feedback Reporting in Fixed and Mobile Networks". Proceedings of international scientific conference ICWMC2007(the third International Conference on Wireless and Mobile Communications), March 2007, pp. 1 – 6, ISBN: 0-7695-2796-5, Guadeloupe, 2007
- [15] D. Komosny and V. Novotny, "Tree structure for source-specific multicast with feedback aggregation," 2007, pp. 0-7695.
- [16] R. Burget, D. Komosny, and M. Simek, "Transmitting Hierarchical Aggregation Information Using RTCP Protocol," IJCSNS, vol. 7, pp. 11-44, 2007.
- [17] V. Novotn and D. Komosn, "Large-scale RTCP feedback optimization," Journal of Networks, vol. 3, p. 1, 2008.

- [18] S. Shahbazi, K. Jumari, and M. Ismail, "A new design for improvement of scalable-RTCP," 2009, pp. 594-598.
- [19] S. Shahbazi, K. Jumari, and M. Ismail, "Error Resistant Real-Time Transport Control Protocol," Am. J. Engg. & Applied Sci, vol. 2, pp. 620-627, 2009.

AUTHORS PROFILE



Adel Nadhem Naeem: He completed his bachelor degree in computer science in 2004 from Shat Al-Arab University College, worked as IT in Iraqna Telecommunications Company from October 2004 until June 2007, started his master's degree in July 2007 in computer science school, USM, and completed it in August 2008, now he is a PhD candidate in National

Advance IPv6, USM. He is one of the researchers that help to develop and improve the communications and networking field, working in multimedia conferencing area.



Ali Abdulqader Bin Salem: received B.S (computer science) degree from Al-Ahgaff University, Yemen in 2006 and M.S (computer science) from University Science Malaysia (USM), Malaysia in 2009. Currently, he is a PhD student at National Advance IPv6 Center (NAv6), (USM). His current research interests include wireless LAN, multimedia QoS, and video

transmission over wireless, distributed system, P2P, and client-server architecture.



Mohammed Faiz Aboalmaaly: A PhD candidate, He received his bachelor degree in software engineering from Mansour University College (IRAQ) and a master's degree in computer science from Univeriti Sains Malaysia (Malaysia). His PhD. research is mainly focused on Overlay Networks. He is interested

in several areas of research such as Multimedia Conferencing, Mobile Ad-hoc Network (MANET) and Parallel Computing.



Professor Dr. SureswaranRamadass: is a Professor and the Director of the National Advanced IPv6 Centre of Excellence (NAV6) at UniversitiSains Malaysia.

Dr. Sureswaran obtained his BsEE/CE (Magna Cum Laude) and Masters in Electrical and Computer Engineering from the University of Miami in 1987 and 1990 respectively. He obtained his PhD from

UniversitiSains Malaysia (USM) in 2000 while serving as a full time faculty in the School of Computer Sciences.

Dr. Sureswaran's recent achievements include being awarded the AnugerahTokoh Negara (National Academic Leader) for Innovation and Commercialization in 2008 by the Minister of Science and Technology. He was also awarded the Malaysian Innovation Award by the Prime Minister in 2007. Dr. Sureswaran is also the founder and headed the team that successfully took Mlabs Systems Berhad, a high technology video conferencing company to a successful listing on the Malaysian Stock Exchange in 2005. Mlabs is the first, and so far, only university based company to be listed in Malaysia.

PERFORMANCE ANALYSIS OF NONLINEAR DISTORTIONS FOR DOWNLINK MC-CDMA SYSTEMS

Labib Francis Gergis

Misr Academy for Engineering and Technology Mansoura, Egypt

Abstract-Multi-carrier (MC) scheme became a promising technique for its spectral efficiency robustness against frequency-selective fading. Multi-carrier code division multiple access (MC-CDMA) is a powerful modulation technique that is being considered in many emerging broadband communication systems. MC-CDMA combines the advantages of multicarrier modulation with that of code-division multiple access (CDMA) to offer reliable highdata-rate downlink cellular communication The MC-CDMA services. signals superposition of many narrow-band signals and, as a result suffer from strong envelope fluctuations which make them very prone to nonlinear effects introduced by high power amplifier (HPA). HPA introduces conversion in both amplitude and phase. In this paper we have focused on the signals at the output of the nonlinear distorting device. A practical technique for determining the bit error rate (BER) of downlink MC-CDMA systems using binary phase- shift keying (BPSK) modulation scheme. The results are applicable to systems employing a coherent demodulation maximal ratio combining (MRC) and equal gain combining (EGC).

Keywords- MC-CDMA systems, high power amplifiers, nonlinear distortions, maximal ratio combining (*MRC*), equal gain combining (*EGC*).

1. INTRODUCTION

Future wireless radio networks need to make efficient use of the frequency spectrum by providing high capacity in terms of number of users allowed in the system. Due to the advantages of spectrum efficiency, interference immunity, high data rate, and sensitivity to selective fading channels. Multi-carrier Coded-division multiple-access (MC-CDMA) appears to be a recommended candidate for future radio communication systems. It exploits the advantages of spread spectrum and the advantages of multi-carrier systems [1].

MC-CDMA signals are considered as superposition of many narrow-band signals, and as a result suffer from strong envelope fluctuations which make them very prone to nonlinear effects introduced by high power amplifiers (*HPA's*) [2].

Power amplifiers (PA's) are components in many communication system. The linearity of a PA response constitutes an important factor that ensures signal integrity and reliable performance of the communication system. High power amplifiers in microwave range suffer from the effects of amplitude modulation to amplitude modulation distortion (AM/AM), and amplitude modulation to phase modulation distortion (AM/PM) [3], during conversions caused by the HPA amplifiers. These distortions can cause intermodulation (IM) distortion, which is undesirable to system designs. The effects of AM/AM and AM/PM degrade distortions the bit error performance of a communication channel.

The amplitude and phase modulation distortions are minimized using linearization method. The linearization method requires modeling the characteristics of the amplitude distortion and phase distortion of the HPA. A Saleh model [4] for traveling wave tube (TWT) amplifiers, has been used to provide the linearization method and applied to measured data from HPA that characterize

the distortion caused by the HPA. The measured data provides a performance curve indicating nonlinear distortion. The forward Saleh model is a mathematical equation that describes the amplitude and phase modulation distortions of the HPA.

The BER analysis of MC-CDMA based on considering different kinds of assumptions, so far, have been dedicated in numerous researches in advance.

Performance enhancement of MC-CDMA system through, space time trellis code (STTC) site diversity with multiple input multiple output (MIMO) technique was introduced in [5].

A method efficiently suppressing multiple access interferences (MAI) in MC-CDMA to improve the system capacity was proposed in [6].

The performance of fully loaded downlink MC-CDMA systems in the presence of residual frequency offset (RFO) in multipath Rayleigh fading channels with minimum mean square error (MMSE) equalizers was presented in [7].

The performance analysis of MC-CDMA communication systems over Nakagami-m fading channels was considered in [8].

A downlink MC-CDMA system using binary phase-shift keying (BPSK) modulation scheme and maximal ratio combining (MRC) in frequency-selective Rician fading channels was illustrated in [9].

The aim of this paper is to analyze the influences of the effects of the nonlinear distortions introduced by HPA in downlink MC-CDMA over Rayleigh fading channel for mobile satellite communication systems. The structure of this paper is as follows. The basic principles model of transmitter system is presented and described in more details in section 2. Section 3 summarizes the HPA baseband models, which is most commonly used in mobile satellite communication systems. Subsequently in section 4, the channel model is described. The receiver model will be described in section Performance analysis of linearized downlink MC-CDMA based signal is carried out for both EGC and MRC.

2. MC-CDMA TRANSMITTER MODEL

The input data symbols, $a_m[k]$, are assumed to be binary antipodal where k denotes the kth bit interval and m denotes the mth user. It is assumed that $a_m[k]$ takes on values of -1 and +1 with equal probability.

As shown in Figure. 1, a single data symbol is replicated into N parallel copies. Each branch of the parallel stream is multiplied by a chip from a spreading code of length N. Each copy is then binary phase-shift keying (BPSK) modulated to a subcarrier spaced apart from its neighboring subcarriers by F/T_b Hz where F is an integer number. An MC-CDMA signal consists of the sum of the outputs of these branches.

As illustrated in Figure. 1, the transmitted signal for MC-CDMA system corresponding to the *kth* data bit of the *mth* user is [10]

$$S_{m}(t) = \sum_{i=0}^{N-1} C_{m}[i] \ a_{m} \ [k] \cdot \\ cos (2\pi f_{c}t + 2\pi i \ (F/T_{b})t \cdot \\ P_{T_{b}} \ (t-kT_{b}) \\ C_{m}[i] \in \{-1,1\}$$
 (1)

where $C_m[0]$, $C_m[1]$,, $C_m[N-1]$ represent the spreading code of the *mth* user and P_{Tb} (t) is an unit amplitude pulse that is non-zero in the interval $[0,T_b]$.

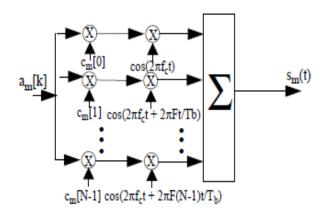


Fig. 1 Transmitter Model of MC-CDMA System

3. NONLINEARITY EFFECTS ON MC-CDMA SIGNAL

The response of broadband power amplifiers can have precarious memory effect. The influence of a memory-less nonlinearity U(.) can be decomposed into an amplitude distortion (AM/AM) and a phase distortion (AM/PM), which are both functions of the amplitude of the input signal to HPA. The complex signal $S_o(t)$ at the output of HPA, can be defined as [11]

$$S_{o}(t) = U\{S_{m}(t)\} = A(|S_{m}(t)|).$$

$$\exp(j \Phi(|S_{m}(t)|)) S_{m}(t) \qquad (2)$$

 $A[S_m(t)]$ and $\Phi[S_m(t)]$ are the corresponding AM/AM and AM/PM characteristics respectively, both dependent exclusively on U_x , which is the input modulus to HPA, they are defined as Saleh Model for HPA [12]:

$$A[U_x] = \alpha_a \ U_x \ / \ 1 + \beta_a \ U_x^2$$

$$\Phi[U_x] = \alpha_{\Phi} \ U_x \ / \ 1 + \beta_{\Phi} \ U_x^2$$
(3)

The values of α_a , β_a , α_{Φ} and β_{Φ} are defined in [3].

The corresponding AM/AM and AM/PM curves so scaled are depicted in Fig. 2.

While for solid state power amplifier types (SSPA's) AM/AM and AM/PM can be defined as

$$A[U_x] = U_x / [1 + (U_x / A_{max})^{2p}]^{1/2p}$$

$$\Phi[U_x] = 0$$
(4)

 A_{max} is the maximum output amplitude, and p is a constant controls the smoothness of the transition.

$$A_{max} = \max (A[U_x]) = \alpha_a A_s / 2$$
 (5)

where A_s is the input saturation amplitude equals $1/\sqrt{\beta_a}$

The HPA operation in the region of its nonlinear characteristic causes a nonlinear distortion of a transmitted signal, that subsequently results in increasing the bit error rate (BER), and the out-of-band energy radiation (spectral spreading).

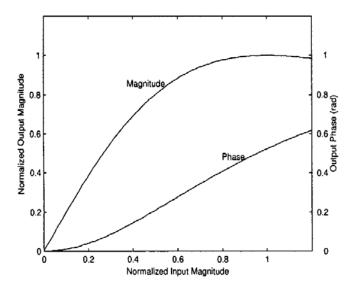


Fig. 2. AM/AM and AM/PM characteristics of the Saleh model For TWTA HPA's

The operating point of HPA is defined by input back-off (*IBO*) parameter which corresponds to the ratio of saturated output power (P_o), and the average input power (P_{av}) [13]:

$$IBO_{dB} = 10 \log_{10} (P_o / P_{av})$$
 (6)

The measure of effects due to the nonlinear HPA could be decreased by the selection of relatively high values of *IBO*

The output of HPA defined in Fig. 3, is expressed as

$$b_{y} = A [U_{x}] e^{j(\alpha x + \Phi[Ux])}$$
 (7)

where the input-output functional relation of the HPA has been defined as a *transfer function*. Hence in order to obtain linearization, it may be necessary to estimate a discrete inverse multiplicative function HPA⁻¹ [.] such that

$$b_x = b_y$$
 . HPA⁻¹ [U_y] (8)

An alternative expression for the AM/AM distortion in (7), convenient for the theoretical formulation of the linearizer, is obtained by

multiplying the saturation input amplitude As in the expression (3). This gives

$$A[U_x] = (A^2_s \alpha_a U_x) / (A^2_s + A^2_s \beta_a U^2_x)$$

$$A[U_x] = (A^2_s \alpha_a U_x) / (A^2_s + U^2_x)$$
 (9)

The theoretical AM/AM inverse transfer function $A^{-1}[.]$ could be determined by solving (9) for $U_x = A \{ A^{-1} [U_x] \}$

$$[u] = (A^{2}_{s} \alpha_{a} / 2U) \cdot \frac{1 - (2U / A_{s} \alpha_{a})^{2}}{1 - (2U / A_{s} \alpha_{a})^{2}}$$
(10)

Considering the alternative configurations shown in Fig. 3, where the same input-output function is applied as a pre-distorter [PD] for the linearization of the same HPA. Letting $\psi[.]$ denote the AM/PM characteristic of the PD block.

For the case of a Pre-distortion, we have [12]:

$$b_{pout} = A^{-I} [U_x] e^{j(\alpha x + \psi[Ux])}$$

$$b_y = A [A^{-I} [U_x]] \cdot$$

$$e^{j(\alpha x + \psi[Ux] + \Phi[A-I[Ux])}$$
(12)

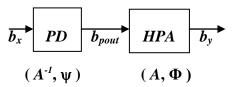


Fig. 3. Pre-distortion for HPA Linearization

The ideal AM/PM correction requires that

$$\psi[U_x] = -\Phi \{ A^{-1}[U_x] \}$$
 (13)

$$b_{pin} = A \left[U_x \right] e^{j(\alpha x + \Phi[Ux])}$$
 (14)

$$b_{y} = A^{-1} \{ A [U_{x}] \}$$

$$e^{j(\alpha x + \Phi[Ux] + \psi[A-I[Ux])}$$
(15)

Pre-distortion linearization idea, as depicted in Fig. 4, can be used to linearize over a wide

bandwidth. This is achieved by pre-distortion of the signal prior to amplification with the inverse characteristics of the distortion that will be imposed by the power amplifier. Thus the output of the HPA is a linear function of the input to the predistorter.

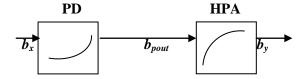


Fig. 4. Basic System Functional Diagram of Predistortion Linearization

A description of the ideal theoretic AM/AM and AM/PM inverse characteristics, valid for the normalized Saleh's HPA model is shown in Fig. 5

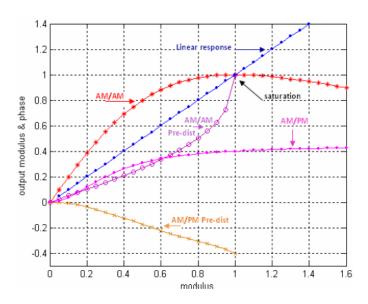


Fig. 5. AM/AM and AM/PM pre-distortion for the Saleh model

4. CHANNEL MODEL

A frequency-selective fading channel with $1/T_b << BW_c << F/T_b$ is considered, where BW_c is the coherence bandwidth. Each modulated subcarrier with transmission bandwidth of $1/T_b$ does not experience significant dispersion $(T_b >> T_d)$. Doppler shifts are very small, it is also assumed that the amplitude and phase remain constant over the symbol duration, T_b .

For downlink transmissions, a terminal receives interfering signal designated for other users (m = 1, 2,, M-1) through the same channel as the wanted signal (m=0), the transfer function of the continuous-time fading channel for all transmissions from the base station to user m = 0 can be represented as

$$H(f_c + i F/T_b) = \rho_{m,i} e^{j\theta_{m,i}}$$
 (16)

where $\rho_{m,i}$, and $\theta_{m,i}$, are the random amplitude and phase of the channel of the mth user at frequency $f_c + i$ (F/T_b). $\rho_{m,i}$ are assumed to be independent and identically distributed (IID) Rayleigh random variables. The random phases, $\theta_{m,i}$ are assumed to be IID random variables uniform on the interval of $\{0, 2\pi\}$ for all users and subcarriers.

5. RECEIVER MODEL

For M active transmitters, the received signal is [10]

$$r(t) = \sum_{m=0}^{M-1} \sum_{i=0}^{N-1} \rho_{m,i} C_m[i] a_m[k] \cdot \cos(2\pi f_c t + 2\pi i [F/T_b]t + \theta_{m,i}) + n(t)$$
(17)

where n(t) is additive white Gaussian noise (AWGN). The local-mean power at the *ith* subcarrier of the mth user is defined to be $\rho_{m,i} = E\overline{\rho^2}_{m,i} / 2$. Assuming the local-mean powers of the subcarriers are equal, the total local-mean power of the mth user is equal to $\overline{p_m} = N \, \overline{p_{m,i}}$.

As shown in Figure. 6, the first step in obtaining the *decision* variable involves demodulating each of subcarriers of the received signal, which includes applying a phase correction, θ_i , and multiplying the *ith* subcarrier signal by a gain correction, d_i .

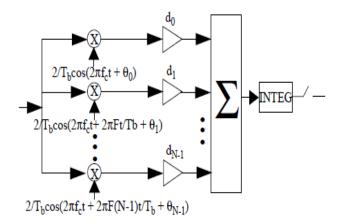


Fig. 6 Receiver Model

After adding the subcarrier signals together, the combined signal is then integrated and sampled to yield decision, V_o . For the kth bit, the decision variable is

$$V_{o} = \sum_{m=0}^{M-1} \sum_{i=0}^{N-1} \rho_{m,i} C_{m}[i] d_{i} a_{m} [k] \cdot \int_{kT_{b}}^{(k+1)T_{b}} cos (2\pi f_{c}t + 2\pi F [i/T_{b}]t + \theta_{m,i}) \cdot cos (2\pi f_{c}t + 2\pi F [i/T_{b}]t + \theta_{m,i}) dt + \eta$$
(18)

where the corresponding AWGN term, η , is given as

$$\eta = \sum_{i=0}^{N-I} \int_{kT_b}^{(k+I)T_b} n(t) (2/T_b) d_i \cdot \cos (2\pi f_c t + 2\pi F [i/T_b]t + \theta_{m,i}) dt \quad (19)$$

Considering the two standard diversity reception techniques: Equal Gain Combining (EGC) and Maximum Ratio Combining (MRC)

With EGC, the gain correction factor at the *ith* subcarrier is given as

$$d_{0,i} = c_0 [i] \tag{20}$$

This scheme yields the decision variable

$$V_{o} = a_{o} [k] \sum_{i=0}^{N-1} \rho_{,i0} + \beta_{int} + \eta$$
 (21)

 $a_o[k] \sum_{i=0}^{N-1} \rho_{,i\theta}$ represents the desired signal, and

the interference term, β_{int} , is defined by

$$\beta_{int} = \sum_{m=0}^{M-1} \sum_{i=0}^{N-1} C_m[i] a_m [k] C_0[i] \rho_{m,i} \cos \theta_{m,i}$$
(22)

For MRC scheme, the gain correction factor at the *ith* subcarrier is given as

$$d_{\theta,i} = \rho_{\theta,i} c_{\theta}[i] \tag{23}$$

The decision variable for MRC scheme is expressed as

$$V_o = a_o [k] \sum_{i=0}^{N-1} \rho^2_{,i0} + \beta_{int} + \eta$$
 (24)

where, the interference term, β_{int} , is defined in this case by

$$\beta_{int} = \sum_{m=0}^{M-1} \sum_{i=0}^{N-1} C_m[i] a_m [k] C_0[i] \rho_{m,i} \rho_{o,i} \cos \theta_{m,i}$$
(25)

6. PERFORMANCE ANALYSIS

The downlink BER had been calculated as [10] 1-with EGC

$$BER =$$

1/2 erfc
$$\sqrt{\frac{\pi}{4 \frac{2(M-1)}{N}}} (26)$$

2- with MRC

BER =
$$\frac{P_{\theta} T_{b}}{P_{\theta} T_{b}} = \frac{1/2 \text{ erfc}}{\sqrt{\frac{2(M-1)}{N} - \frac{-1}{(1-\pi/4)P_{\theta}T_{b} + N_{\theta}}}} (27)$$

7. NUMERICAL RESULTS

A fair measure is given by using the normalized minimal signal-to-noise ratio

$$SNR_o = 10 \log (P_o T_b / N_o) (dB)$$
 (28)

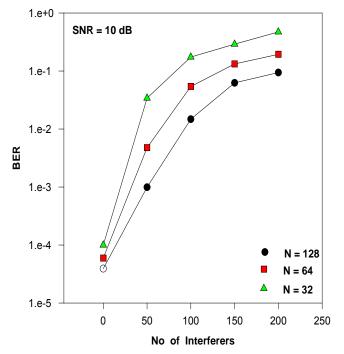
which is needed to achieve the wanted BER. T_b is the equivalent duration for one information bit, N_o is the two sided spectral noise density, and P_o is the given reference power of HPA. The SNR_o can be minimized by optimization of the HPA backoff. This becomes more clear, when eq. (6) is used in eq. (28):

$$SNR_o = 10 \log (P_o T_b P_{av} / N_o P_{av})$$

= $10 \log (E_b / N_o) + OBO$ (29)

The average downlink bit error rate (BER) versus the number of interferes are examined. For the sake of comparison, the BER for both types of diversity, EGC and MRC are illustrated under interferers numbers, $N=32,\,64,\,$ and 128, with SNR=10 dB in Figures 7, and 8.

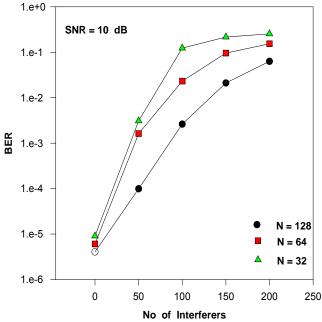
It can be seen that for a small numbers of users, MRC outperforms EGC. It was also demonstrated the PD effect to mitigate the nonlinearity distortions introduced from HPA in Fig. 9, and Fig. 10.



1.e+0 EGC case 1.e-1 N = 1281.e-2 1.e-3 1.e-4 1.e-5 1.e-6 1.e-7 1.e-8 1.e-9 with PD (m=0 interferers) 1.e-10 without PD (oBo=5 dB) (m=0 interferers) 1.e-11 with PD (m =70 interferers) 1.e-12 without PD (oBo = 5 dB) (m = 70 interferers) 1.e-13 10 0 5 15 20 25 30 35 SNR dB

Fig. 7. BER versus the No. of Interferers for EGC case

Fig. 9. BER versus the SNR using PD for EGC case



No of Interferers

Fig. 8. BER versus the No. of Interferers for MRC case

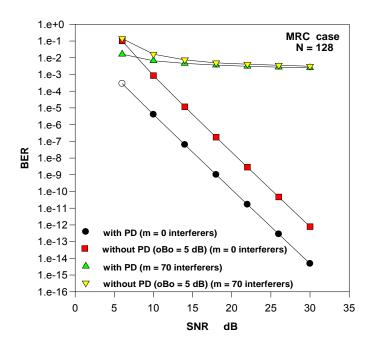


Fig. 10. BER versus the SNR using PD for MRC case

8. CONCLUSIONS

In this paper, the downlink transmission in MC-CDMA systems with nonlinear HPA of transmitter over frequency-selective fading channels was considered. This paper presented results on a novel modulation, diversity, and multiple access technique.

For two diversity techniques considered, MRC performed better than EGC.

The performance of MC-CDMA would be affected by nonlinearities introduced from HPA's in the transmitter.

From previous discussions and plotted results, it can be concluded that in order to reduce the sensitivity of a MC-CDMA system to the nonlinear amplification, it is recommended to choose a special technique to mitigate these distortions. PD schemes had been selected to do this mission, they attended to achieve a significantly improve overall system performance.

REFERENCES

- [1] J. Iong, and Z. Chen, "Theoretical Analysis of an MC-CDMA System in Cellular Environments" Journal of Science and Engineering Technology, Vol. 6, No. 1, pp. 37-50, 2010.
- [2] M. Frikel, S. Safi, B. Targui, and M. Saad," Channel Identification Using Chaos for an Uplink/Downlink Multicarrier Code Division Multiple Access System", JOURNAL OF TELECOMMUNICATIONS AND INFORMATION TECHNOLOGY, Vol. 1, pp. 48-54, 2010.
- [3] M. Chen, and O. Collins, "Multilevel Coding for Nonlinear ISI Channels", IEEE TRANSACTIONS ON INFORMATION THEORY. VOL. 55, NO. 5, MAY 2009.
- [4] J. Gazda, P. Drotar, D. Kocur, and P. Galajda," Performance Improvements of MC-CDMA Microstatistic Multi-User Detection in Nonlinear Fading Channels Using Spreading Code Selection," TUBITAK. Turk J Elec Engin. VOL. 8, NO. 1, April 2009.

- [5] N. Kumaratharan, S. Jayapriya, and P. Dananjayan, "Performance Enhancement of MC-CDMA System through STBC based STTC Site Diversity", International Journal of Computer and Electrical Engineering, Vol. 2, No. 1, February, 2010.
- [6] B. Ness, "EQUAL BER POWER CONTROL FOR UPLINK MC-CDMA WITH MMSE SUCCESSIVE INTERFERENCE CANCELLATION," Patent No. US 7,620,096 B2, Nov. 2009.
- [7] P. Reddy, and V. Reddy," BER Degradation of MC-CDMA at high SNR with MMSE Equalization and Residual Frequency Offset," EURASIP Journal on Wireless Communications and Networking, Volume 2009, Article ID 293264, 2009.
- [8] J. Iong, and Z. Chen, "PERFORMANCE ANALYSIS OF MC-CDMA COMMUNICATION SYSTEMS OVER NAKAGAMI-M ENVIRONMENTS", Journal of Marine Science and Technology, Vol. 14, No. 1, pp. 58-63, 2006.
- [9] Z. Hou, and V. dubey, "BER Performance for Downlink MC-CDMA Systems over Rician Fading Channels", EURASIP Journal on Applier Signal Processing, pp. 709-717, 2005.
- [10] N. Yee, J. Linnartz, and G. Fettweis,"
 MULTICARRER CDMA IN INDOOR
 WIRELESS RADIO NETWORKS", The
 Fourth International Symposium on
 Personal Indoor and Mobile Radio
 Communications, PACIFICO
 YOKOHAMA, JAPAN, September. 1993.
- [11] A. Perotti, P. Rrmlein, and S. Benedetto, "
 Adaptive Coded Continuous-Phase
 Modulations for Frequency-Division
 Multiuser Systems", ADVANCES IN
 ELECTRONICS AND
 TELECOMMUNICATIONS, VOL. 1, NO.
 1, APRIL 2010
- [12] S. Chang, "An efficient compensation of TWTA's nonlinear distortion in wideband OFDM systems", IEICE Electronics Express, Vol. 6, No. 2, pp. 111-116, 2009.
- [13] T.Tan," POWER AMPLIFIER
 MODELING AND POWER AMPLIFIER
 PREDISTORTION IN OFDM SYSTEM",
 Journal of Science & Technology
 Development, Vol 11, No.02 2008

Channel Estimation Algorithms, Complexities and LTE Implementation Challenges

Md. Masud Rana
Department of Electronics and Communication Engineering
Khulna University of Engineering and Technology
Khunla, Bangladesh

Abstract—The main purposes of the long term evolution (LTE) are substantially improved enduser throughputs, low latency, reduced user equipment (UE) complexity, high data rate, and significantly improved user experience with full mobility. LTE uses single carrier-frequency division multiple access (SC-FDMA) for uplink transmission and orthogonal frequency division multiple access (OFDMA) for downlink transmission. The major challenges for LTE terminal implementation are efficient channel estimation (CE) method as well as equalization. This paper discusses the basic CE techniques and future direction for research in CE fields. Simulation results demonstraters that the linear mean square error (LMMSE) CE method outperforms the least square (LS) CE method in term of mean square error (MSE) by more than around 3dB. Hence, based on a given LTE systems resources and specifications, a appropriate method among the presented methods can be applied for OFDMA systems.

Keywords-LS, LMMSE, LTE, OFDMA.

I. INTRODUCTION

The wireless evolution has been stimulated by an explosive growing demand for a wide variety of high quality of services in voice, video, and data. This rigorous demand has made an impact on current and future wireless applications, such as digital audio/video broadcasting, wireless local area networks (WLANs), worldwide interoperability for microwave access (WiMAX), wireless fidelity (WiFi), cognitive radio, and 3rd generation partnership project (3GPP) long term evolution (LTE) [1], [2]. LTE uses single carrierfrequency division multiple access (SC-FDMA) for uplink transmission and orthogonal frequency division multiple access (OFDMA) for downlink transmission [3], [4]. SC-FDMA utilizes single carrier modulation and frequency domain equalization, and has similar performance and essentially the same overall complexity as those of OFDMA system. These advanced applications in which the transmitted signal disperses over the time and the frequency domains, show the need for highlydeveloped signal processing algorithms. In particular, one of the main challenges in the mobile communication is a wireless channel that suffers from numerous physical impairments due to multipath propagation, interference from other users or layers, and the time selectivity of a channel [5-9].

Many CE techniques have already been proposed for the LTE OFDMA systems. The simple least square (LS) algorithm, which is independent of the channel model, is commonly used in CE [10-14]. But the radio channel is time-variant; hence a method has to be found in order to perform estimation in a time-varying channel. The minimum mean-squared error (MMSE) estimate has been shown to be better than the LS estimate for CE in wireless communication systems [15]. The important problem of the MMSE estimate is its high computational complexity, which grows exponentially with inspection samples [16]. In [17], a low rank approximation is applied to a linear MMSE (LMMSE) estimator that employs the correlations of the channel. To further improve the system performance, Wiener estimation has been investigated [18]. Although it exhibits the best performance among the existing linear algorithms, it requires accurate knowledge of second order channel statistics, which is not always feasible at a mobile receiver. Also, this scheme requires higher complexity.

This paper outlines the developments of the LTE OFDMA systems, and highlights some upcoming challenges, where advanced signal processing could play a important role in resolving them. Specifically, we investigates various types of CE techniques such as LS, and LMMSE CE methods and find out which is the more efficient one. The performance is measured in terms computational complexity, and mean square error (MSE). Simulation results shows that the LMMSE CE algorithm outperforms the existing LS CE in term of MSE by more than around 3dB. Hence, based on a given LTE systems resources and specifications, a appropriate method among the presented methods can be applied.

The rest of the paper is organized as follows. We give a brief overview of the wireless communication systems in section II. The classification of CE is described in section III. The LS and LMMSE CE methods are describes in section IV and its performance are analyzed in section V. In section VI, we highlight the challenges for LTE terminal implementation. Finally, some conclusions are made in section VII.

The following notations are used in this paper: bold face lower and upper case letters are used to represent vectors and matrices respectively. Superscripts \mathbf{X}^T

 \mathbf{X}^+ denote the transpose and congugate transpose of the \mathbf{X} , and \mathbf{I} is the identity matrix.

II. COMMUNICATION SYSTEMS

Nowadays, cellular mobile phones have become an important tool and part of daily life. In the last decade, cellular systems have experienced fast development and there are currently about two billion users over the world [6]. Mobile penetration is based on population, pay TV and broadband is households.

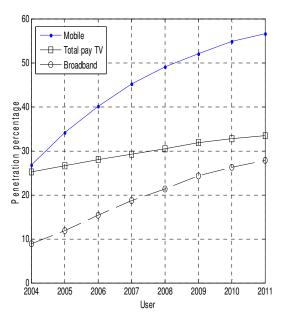


Fig. 1 Mobile is the key growth platform.

The idea of cellular mobile communications is to divide large zones into small cells, and it can provide radio coverage over a wider area than the area of one cell. This concept was developed by researchers at AT & T Bell laboratories during the 1950s and 1960s. The initial cellular system was created by Nippon telephone & telegraph (NTT) in Japan, 1979. From then on, the cellular mobile communication has evolved.

The mobile communication systems are frequently classified as different generations depending of the service offered. The first generation (1G) comprises the analog communication techniques, and it was mainly built on frequency modulation (FM) and frequency multiple division accesses (FDMA). Digital communication techniques appeared in the second generation (2G) systems, and main access schemes are time division multiple access (TDMA) and code division multiple access (CDMA). The two most commonly accepted 2G systems are global system for mobile (GSM) and interim standard-95 (IS-95). These systems mostly offer speech communication, but also data communication limited to rather low transmission rates [7]. The concept of the third generation (3G) system started operations on October, 2002 in Japan.

The 3GPP members started a feasibility study on the enhancement of the universal terrestrial radio access (UTRA) in December 2004, to improve the mobile phone standard to cope with future requirements. This project was called LTE [5], [8]. 3GPP LTE uses SC-FDMA for uplink transmission and OFDMA for downlink transmission [9]. Fig. 2 summarizes the cellular mobile communication systems and its access schemes [10].

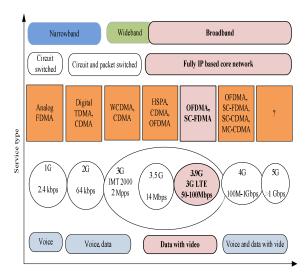


Fig.2 (a): Evolution path in mobile communication systems.

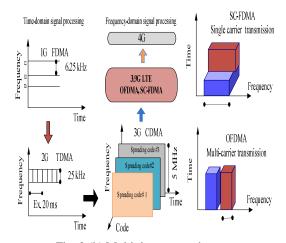


Fig. 2 (b) Multiple access schemes.

From the beginning wireless communications there is a high demand for realistic mobile fading channels. The motive for this significance is that efficient channel models are necessary for the investigation, design, and deployment of wireless communication system for reliable transfer of information between two parties. Correct channel models are also important for testing, parameter optimization, and performance evolution of wireless communication systems. The performance and complexity of signal processing algorithms, transceiver

designs, and smart antennas, employed in future wireless communication systems, are highly dependent on design methods used to model mobile fading channels. The effect of the channel on the transmitted information must be estimated in order to recover the transmitted information correctly [11].

III. CLASSIFICATION OF CE

Channel can be described everything from the source to the destination of a radio signal. This includes the physical medium between the transmitter and the receiver through which the radio signal propagates. On the other hand, CE is the process of characterizing the effect of the physical channel on the input sequence. It can be employed for the purpose of detecting received signal, improve signal to noise ratio (SNR), channel equalization, reduced ISI, mobile localization, and improved system performance [8], [9]. In general, both iterative and noniterative CE techniques can be divided into three categories such as the training CE, blind CE, and semi-blind CE [10], [21].

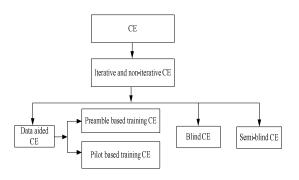


Fig. 3 Outline of the CE.

A. Taining CE

The training CE algorithm requires probe sequences; the receiver can use this probe sequence to reconstruct the transmitted waveform [10]. It has the advantage of being used in any radio communications system quite easily. Even if this is the most popular CE method, it still has its drawbacks. The drawback of training sequence methods is that the probe sequence occupies valuable bandwidth, reducing the throughput of the communication system. This scheme also suffers due to the fact that most communication systems send information lumped frames. It is only after the receipt of the entire frame that the channel estimate can be reconstructed from the embedded probe sequence. Since, the coherence time of the channel might be smaller than the frame time, for rapid fading channels this CE might not be sufficient. Training symbols can be placed either at the beginning of each burst as a preamble or regularly through the burst [21]. Training sequences are

transmitted at certain positions of the OFDMA frequency time pattern, in its place of data as shown in Fig. 4. An amount of training sequences is raise the accuracy of CE, but it is reduces the system efficiency, because there isn't any new information in the training symbols.

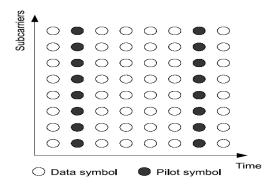


Fig. 4 Positions of data and pilot symbols.

B. Blind CE

A blind CE method requires no training sequences [13]. They exploit certain underlying mathematical information regarding the type of data being transmitted. These CE methods might be bandwidth efficient but still have their own downfalls. These methods are enormously computationally intensive and convergence is slow [21]. A popular category of blind CE method is decision directed algorithms. These methods rely upon the demodulated and detected signal at the receiver to reform the transmitted signal. The drawback of these CE algorithm is that a bit error at the receiver is cause the construction of an erroneous transmitted sequence.

C. Semi-blind CE

Semi-blind CE methods are used a combination of data aided and blind methods [11]. Since, there are a large number of channel coefficients, a large number of pilot symbols may be required. It would result in a decrease of data throughput. To avoid it, the semi-blind CE methods with fewer pilot symbols can be used. As a result, improve system performance in compared with using equal pilots in LS method. Moreover, there is a trend to use superimposition of pilot and data symbols. In fact, these methods by superimposing pilot and data symbols in the same time economize the system bandwidth. But in superimposed training sequence scheme, there is disadvantage due to the interference of information data. So, an accurate CE has been one of the most important issues for reliable mobile communication systems. So, CE can be performed by many ways inserting pilot tones into each OFDMA symbol with a specific period or blind CE.

IV. LS and LMMSE CE ALGORITHMS

Pilot estimators are often achieved by multiplexing training sequence into the data sequence. These pilot symbols allow the receiver to extract channel attenuations and phase rotation estimates for each received symbol, facilitating the compensation of channel fading envelope and phase. A general CE procedure for communication system is shown in Fig. 5.

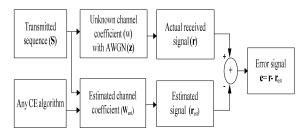


Fig. 5 General CE procedure.

The signal S is transmitted via a unknown time-varying channel \mathbf{w} , and corrupted by an additive white Gaussian noise (AWGN) \mathbf{z} , before being detected in a receiver. The channel coefficient \mathbf{W}_{est} , is estimated using any kind of CE method. In the channel estimator, transmitted signal S is convolved with an estimate of the channel \mathbf{W}_{est} . The error between the received signal and its estimate is

$$\mathbf{e} = \mathbf{r} - \mathbf{r}_{est} \tag{1}$$

The aim of most CE algorithms is to minimize the MSE, while utilizing as little computational resources as possible in the estimation process.

The idea behind LS CE method is to fit a model to measurements in such a way that weighted errors between the estimation and the true model are minimized [14]. The received signal can be written as vector notation as

$$\mathbf{r} = \mathbf{S}\mathbf{w} + \mathbf{z},\tag{2}$$

where $\mathbf{r} = [\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_L]^T$ is the received signal, $\mathbf{S} = \operatorname{diag}[\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_L]$ is the transmitted signal, $\mathbf{w} = [w_1, w_2, \dots, w_L]^T$ is the unknown channel coefficients, and $\mathbf{z} = [z_1, z_2, \dots, z_L]^T$ is additive white Gaussian noise (AWGN). The LS estimate of such a system is obtained by minimizing square distance between the received signal and its estimate as [14]

$$\mathbf{j} = (\mathbf{r} - \mathbf{S}\mathbf{w}) (\mathbf{r} - \mathbf{S}\mathbf{w})^{\dagger}. \tag{3}$$

Now differentiate this with respect to \mathbf{W} and set the results equal to zero to produce [14]:

$$\mathbf{w}_{1S} = (\rho \mathbf{I} + \mathbf{S}\mathbf{S})^{-1} \mathbf{S}^{+} \mathbf{r}, \tag{4}$$

where ρ is regularization parameter and has to be chosen such that the resulting eigenvalues are all defined and the matrix $(\rho \mathbf{I} + \mathbf{S}\mathbf{S})^{-1}$ is least perturbed. Here the channel is considered as a deterministic parameter and no knowledge on its noise statistics is needed. The LS estimator is computationally simple but problem is that the inversion of the square matrix turns out to be ill-conditioned (sometime). So, it will need to regularize the eigenvalues of the matrix to be inverted by adding a small constant term to the diagonal [14].

MMSE CE method proposes at the minimization of the MSE between the actual and estimated channel impulse response (CIR). The most important problem of the MMSE estimate is its high computational complexity, which grows exponentially with inspection samples [15], [16]. In [17], a low rank approximation is applied to a linear MMSE (LMMSE) estimator that employs the correlations of the channel. The general expression of LMMSE is described as

$$\mathbf{w}_{\text{est}} = \mathbf{R}_{\mathbf{ww}} (\mathbf{R}_{\mathbf{ww}} + \Gamma \mathbf{I}/\text{SNR})^{-1} \mathbf{w}_{\text{LS}}, \quad (5)$$

where $\mathbf{R}_{\mathbf{w}\mathbf{w}}$ is the auto-covariance matrix of \mathbf{w} , \mathbf{w}_{LS} is the channel response in LS estimation, and Γ is a constant depending on the modulation constellation

$$\Gamma = \mathbb{E}[|\mathbf{S}_{k}|^{2}] \mathbb{E}[|1/\mathbf{S}_{k}|^{2}]. \tag{6}$$

For QPSK modulation, r is 1[17]. Here, \mathbf{W}_{LS} is not very important issue in the matrix computation, the inversion of \mathbf{R}_{ww} does not require to be estimated every time the transmitted sybmols in \mathbf{W}_{LS} varies. Also, if signal to noise ratio (SNR) and \mathbf{R}_{ww} are identified earlier or are set to fixed nominal values, the matrix $(\mathbf{R}_{ww} + \Gamma \mathbf{I}/SNR)^{-1}$ needs to be computed at once. Under these situation, the estimation requires L multiplications per tone.

In order to calculate computational complexity, we assume that the evaluation of the scalar addition or subtraction needs L addition and multiplying the scalar by the vector requires L multiplications, and multiplying two matrix need 4L multiplications and 4L-1 additions. Table I summarizes the computational complexity of the different CE methods.

Table I Complexity of the CE methods

Operation	LS CE	LMMSE CE
Matrix inversion	1	2
Multiplication	11L	17L
Addition	11L - 3	17L - 5

We calculate the number of complex addition and multiplications which are needed to implement the algorithm. It shows that the LS CE algorithm has lower complexity than LMMSE method. For this LMMSE estimator, the main contribution to the complexity comes from the term $\mathbf{R}_{ww}(\mathbf{R}_{ww} + \Gamma \mathbf{I}/SNR)^{-1}$.

V. ANALYTICAL RESULT

In this simulations, we consider a system operating with a bandwidth of 1.25MHz, with a total symbol period of 520µs, of which 10 µs is a cyclic prefix. The entire channel bandwidth is divided into 128 subcarriers, implemented by 128-point IDFT. Sampling is performed with a 1.92MHz. The data symbol is based on BPSK. In practice, the ideal channel coefficient is unavailable, so estimated channel coefficient must be used instead. The more accurate estimated channel coefficient is, the better MSE performance of the CE will achieve. The performance is measured using MSE between the actual and the estimated channel response. Fig. 6 shows the MSE versus SNR for the different channel estimators. We can see that LMMSE CE can always achieve better performance than LS CE. The main reason is, LMMSE CE method uses channel correlation as well as SNR but the LS CE method does not uses channel correlation. Finally, it concludes that the LMMSE CE method has higher computational complexity and around 3dB better performance compared with the LS CE method.

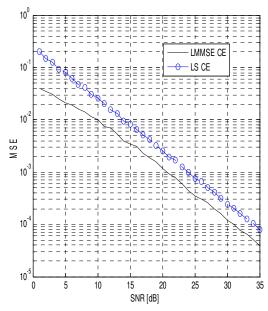


Fig. 6 MSE of the LS and LMMSE CE methods.

VI. IMPLEMENTATION CHALLENGES

LTE meets the important obligations of next generation mobile communications, but still falls short

on some preferred requirements such as cell-edge spectral efficiency in the uplink transmission [19]. LTE implementation poses the following signal processing challenges in terms of performance, cost and power consumption:

- Regrettably, the development of data rates is not matched by advanced in semiconductor structures, terminal power consumption improvements. Therefore, advanced signal processing architectures as well as algorithms are needed to cope with these data rates [19].
- High performance multiple input multiput output (MIMO) receivers such as sphere decoders, maximum likelihood receivers offer substantial system performance gains but enforce an implementation challenge, especially when the high peak data rates are targeted [19].
- LTE utilizes precoding, which requires accurate CE. Advanced methods like iterative decision directed CE and pilot based CE offer system performance improvements, but pose again a computational complexity challenge [19].
- LTE has a large "toolkit" of MIMO methods and adaptive methods. The choice and combination of the accurate technique in a cell with heterogeneous devices, channel conditions and bursty data services is a challenge [19].
- It is very difficult to implement many antennas in a small hand portable unit. In near future, we need to use wearable antenna on head.
- LTE roll-out will be gradual in most casesinterworking with other standards such as GSM or HSPA is required for a long time. This imposes not only a cost and computational complexity issue. One of the reasons many early 3G terminals had poor power consumption was the need for second generation (2G) cell search and handover in addition to normal 3G operation. Reduced talk-time for dual-mode devices is not suitable [19]. Fig. 7 shows the estimated complexity based on the baseline receiver. Note that the complexity of the LTE receiver grows linearly with respect to system bandwidth and the corresponding maximum nominal throughput. Interestingly, MIMO mode requires less than double the SIMO mode complexity.

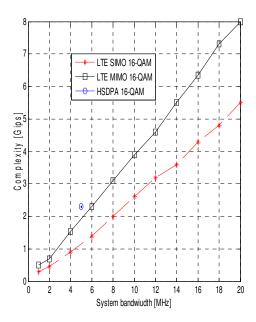


Fig. 7 Complexity of LTE receiver.

VII. CONCLUSION

An accurate CE is one of the most important issues for reliable future wireless communication systems such as LTE. In this paper, we briefly insvesteget LS and LMMSE CE techniques for LTE terminal implemtation. Simulations demonstrated that the MSE performance of the LMMSE CE algorithm is at least 3dB better than existing LS estimator. Even though, the LMMSE CE technique requires a little high computational complexity, the advantage in the MSE and convergence towards true channel coefficient may be significantly useful for future mobile communications which allow broadband multimedia Internet access and wireless connection anywhere, and any time. This paper also discusses the challenges imposed by developments in the LTE terminal implementation. Hence, based on a given LTE systems resources and specifications, a appropriate method among the presented methods can be applied.

REFERENCES

- [1] Q. Li, G. Li, W. Lee, M. I. Lee, D. Mazzarese, B. Clerckx, and Z. Li, "MIMO techniques in WiMAX and LTE: a feature overview," *IEEE Commun. Magazine*, vol. 48, no. 5, pp. 86–92, May 2010.
- [2] M. Alasti, B. Neekzad, C. J. Hui, and R. Vannithamby, "Quality of service in WiMAX and LTE networks," *IEEE Commun. Magazine*, vol. 48, no. 5, pp. 104–111, May 2010.
- [3] Z. Lin, P. Xiao, B. Vucetic, and M. Sellathurai, "Analysis of receiver algorithms for LTE SC-FDMA based uplink MIMO systems," *IEEE Trans. on Wireless Commun.*, vol. 9, no. 1, pp. 60–65, Jan. 2010.
- [4] M. Zhou, B. Jiang, T. Li, W. Zhong, and X. Gao, "DCT-based channel estimation techniques for LTE uplink," In

- Proc. Personal, Indoor and Mobile Radio Commun., Sept. 2009
- [5] L. Sanguinetti, M. Morelli, and H. V. Poor, "Frame detection and timing acquisition for OFDM transmissions with unknown interference," *IEEE Trans. on Wireless Commun.*, vol. 9, no. 3, pp. 1226–1236, Mar. 2010.
- [6] M. Morelli, G. Imbarlina, and M. Moretti, "Estimation of residual carrier and sampling frequency offsets in OFDM-SDMA uplink transmissions," *IEEE Trans. on Wireless Commun.*, vol. 9, no. 2, pp. 734–744, Feb. 2010.
- [7] Z. Wang, Y. Xin, G. Mathew, and X. Wang, "Multipath parameter estimation for radio propagation channel measurements with a vector network analyzer," *IEEE Trans.* on Vehicular Technology, vol. 59, no. 1, pp. 48–52, Mar. 2010.
- [8] J. I. Montojo and L. B. Milstein, "Channel estimation for non-ideal OFDM systems," *IEEE Trans. on Commun.*, vol. 58, no. 1, pp. 146–156, Jan. 2010.
- [9] M. H. Hsieh and C. H. We, "Channel estimation for OFDM systems based on comb-type pilot arrangement in frequency selective fading channels," *IEEE Trans. on Consumer Electronics*, vol. 44, no. 1, pp. 217–225, Feb. 2004.
- [10] F. Wan, W. P. Zhu, and M. N. S. Swamy, "A semiblind channel estimation approach for MIMO-OFDM systems," *IEEE Trans. on Signal Process.*, vol. 56, no. 7, pp. 2821– 2834, July 2008.
- [11] H. Wu, X. Dai, and H. Zhang, "Semi-blind channel estimation for the uplink of multi-carrier code-division multiple access systems with timing offset," *IET Commun.*, vol. 3, no. 12, pp. 1832–1842, 2009.
- [12] B. Karakaya, H. Arslan, and H. A. Curpan, "An adaptive channel interpolator based on Kalman filter for LTE uplink in high Doppler spread environments," *EURASIP Journal* on Wireless Commun. And Networking, vol. 2009, Article ID 893751, 2009.
- [13] T. Fusco and M. Tanda, "Blind synchronization for OFDM systems in multipath channels," *IEEE Trans. on Wireless Commun.*, vol. 8, no. 3, pp. 1340–1348, Mar. 2009.
- [14] A. Ancora, C. B. Meili, and D. T. Slock, "Down sampled impulse response least-squares channel estimation for LTE OFDMA," In Proc. International Conference on Acoustics, Speech, and Signal Processing, pp. III–293–III–296, Apr. 2007.
- J. J. V. D. Beek, O. Edfors, and M. Sandell, "On channel estimation in OFDM systems," *In Proc. Vehicular Technology Conference*, vol. 2, pp. 815–819, Sept. 1995.
 M. H. Hsieh and C. H. We, "Channel estimation for OFDM
- [16] M. H. Hsieh and C. H. We, "Channel estimation for OFDM systems based on comb-type pilot arrangement in frequency selective fading channels," *IEEE Trans. on Consumer Electronics*, vol. 44, no. 1, pp. 217–225, 1998.
- [17] O. Edfors, M. Sandell, J.J. V. D. Beek, S. K. Wilson, and P. O. Borjesson, "OFDM channel estimation by singular value decomposition," *IEEE Trans. on Commun.*, vol. 46, no. 7, pp. 931–939, 1998.
- [18] L. A. M. R. D. Temino, C. N. I. Manchon, C. Rom, T. B. Sorensen, and P. Mogensen, "Iterative channel estimation with robust Wiener filtering in LTE downlink," *In Proc. Vehicular Technology Conference*, pp. 1–5, Sept. 2008.
- [19] R. Irmer and S. Chia, "Signal processing challenges for future wireless communications," In Proc. International Conference on Acoustics, Speech, and Signal Processing, pp. 3625–3628, Apr. 2009.
- [20] M. M. Rana, "Channel Estimation Techniques and LTE Terminal Implementation Challenges," in proc. Int. Con. on Computer Sciences and Convergence Information Technology, Bangladesh, December 22, 2010.
- [21] X. G. Doukopoulos and G. V. Moustakides, "Blind adaptive channel estimation in OFDM systems," *IEEE Trans. on Wireless Commun.*, vol. 5, no. 7, pp. 1716–1725, 2006.

IMPLEMENTATION OF WAVELET AND RBF FOR POWER QUALITY DISTURBANCE CLASSIFICATION

Pramila P¹, Puttamadappa C² and S.Purushothaman³

¹ Department of Electrical & Electronics Engineering, Bangalore Institute Of Technology, Bangalore, India ² Department of Electronics & Communication Engineering, SJB Institute Of Technology, Bangalore, India ³ Sun College Of Engineering & Technology, Sunnagar, Kanyakumari, Tamilnadu, India

Abstract

This paper presents application of wavelet and Radial Basis Function (RBF) for power quality disturbance classification. Features are extracted from the electrical signals by using db wavelets. The features obtained from the wavelet are unique to each type of electrical fault. These features are normalized and given to the RBF. The data required are generated by simulating various faults in the test system. The performance of the proposed method is compared with the existing feature extraction techniques. Simulation results show the effectiveness of the proposed method for power quality disturbance classification.

Keywords wavelets, Radial basis function (RBF), Harmonics, Power quality

1. Introduction

Electrical fault arises due to sudden loads, failure in the electrical circuits, lightning, conducted or radiated low and high frequency phenomena. Due to electrical fault in the system, the power quality deteriorates which is an indication of slow failure of the equipment. Many algorithms have been developed by previous researchers for classification of electrical disturbances. Integrated Fourier linear combiner and fuzzy expert system [1] were used for the classification of transient disturbance waveforms in a power system. S-Transform and two dimensional time-time (TT) transform [2] have implemented for electrical fault identification. Patterns generated by S-Transform and TT transform are unique and hence accuracy of identification is high. An adaptive neural network approach for the estimation of harmonic distortions and power quality in power networks are implemented [3]. A hybrid system to automatically detect, locate and classify disturbances affecting power quality in an electrical power system is presented [4]. Least absolute value (LAV) State estimation algorithm has been used to measure the flicker voltage magnitude [5]. The Simulated Annealing (SA) optimization algorithm has been used for measuring the voltage flicker magnitude, frequency and the harmonics contents of the voltage signal for power quality analysis [6]. An algorithm to detect the fundamental frequency is proposed. It is based on the chirp-z transform (CZT) spectral analysis and is able to

observe all standards in force because of its accuracy and working characteristics [7].

A wavelet-based neural classifier integrating the DWT [8] [9], learning vector quantization (LVQ) neural network, and decision-making scheme to become an actual power disturbance classifier. The classifier employed the DWT coefficients as inputs to multiple LVQ neural networks to train and perform waveform recognition, and use the decisionmaking schemes to classify the transient disturbance type. A novel classifier using a rule-based method and a wavelet packet-based hidden Markov model (HMM) [10]. The rule-based method is employed to classify the time-characterized feature disturbances, while the wavelet packet-based HMM is utilized to categorize the frequency-characterized feature power disturbances. Wavelet-multi-resolution decomposition that combines frequency-domain with time-domain analysis for power disturbance feature extraction is proposed [11]. Extracting the features from the wavelet transform coefficients at different scales as inputs to neural networks for classifying the nonstationary signal type have been proposed [12]. A wavelet norm entropy-based effective feature extraction method for power quality disturbance classification problem has been done by [9]. Extracting the squared wavelet transform coefficients (SWTC) at each scale as inputs to the neural networks for classifying the electrical disturbance is proposed [13] [14] [15]. Multi-wavelet-based neural networks with learning vector quantization network are used for power quality disturbances as a powerful classifier [16]. The DWT coefficients as inputs to a single-layer self-organizing map neural network to train and classify the transient disturbance has been used [17]. Fuzzy ARTMAP, Back propagation algorithm and Radial Basis Function (RBF) network

in combination with S Transform, Wavelet transform and Hilbert Transform (HT) for classifying power faults have been used [18]. DWT coefficients have been used as inputs to a refined neuro-fuzzy network to train and classify the power system disturbance [19]. Continuous wavelet transform (CWT) has been used to estimate the disturbance time duration and the DWT to estimate the disturbance amplitude [20]. The two features thus obtained are then used to classify the transient disturbance type. The authors have claimed HT with RBF gives more fault classification from a set of 6 different types of faults with 3000 signals generated using Matlab.

Algorithms should be developed that can classify the type of harmonics and other electrical disturbances. Due to generation of non-stationary random disturbances, many existing classification algorithms are provided with intelligence using artificial neural networks, fuzzy logic and evolutionary algorithms. The artificial neural networks (ANNs) can classify the electrical disturbances in the presence of noise in the signal. In this work, RBF has been implemented for electrical fault classification.

2. Proposed Methodology

This research work proposes wavelets for feature extraction and RBF for classification of electrical fault. In order to achieve maximum classification, proper data input, optimum topology and correct training of RBF with suitable parameters is a must. A large amount of patterns are generated from different fault conditions. Twelve features are generated for each pattern. The faults considered are instantaneous interruption, instantaneous sag, instantaneous swell, momentary interruption, momentary sag, momentary swell, temporary interruption, temporary sag, temporary swell and harmonics. The features obtained are mean, standard deviation, norm, maximum and minimum of signal for Approximation and Details at the 5th level decomposition. These features are given as inputs for the RBF neural network along with labeling that indicates a fault. Training the RBF gives final weights. The final weights are used to identify a fault during testing. Fig. 1 explains the overall sequence of proposed methodology.

Steps in feature extraction from the input signal

- 1. Input voltage wave form is sampled, $f_s=12800$ Hz.
- 2. Decomposition of the signal by wavelets till 5th level.
- 3. Ten features are obtained from Approximation and Detail in the 5^{th} level

- 4. Input the extracted features into the input layer of features.
- 5. Train the RBF network and store the final weights.
- 6. Test the RBF network with new signal having a fault.

3. Wavelet decomposition

Wavelet is a process of analyzing the signal with scaling and shifting to obtain the details of the signal. It decomposes the signal into high frequencies (Detail) and low frequencies (Approximation). The first time getting Approximation and Detail is called level 1 decomposition (Figure 1). Subsequent decompositions are done. by using Approximation or Detail obtained in level 1, to further get the required information. In every level of decomposition, the number of samples in the signal is reduced to 50% of the total samples to Approximation and remaining 50% of the samples to Detail. During subsequent decomposition, only 50% of the samples (Approximation will be further used for getting Approximation and Detail in the subsequent level decomposition).

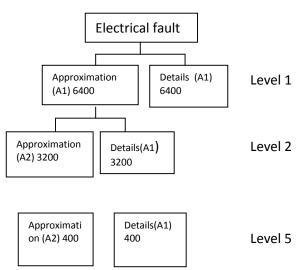


Figure 1 Levels of decomposition

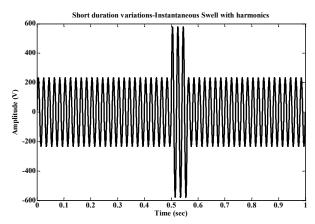


Figure 2 Short duration variations-Instantaneous swell with harmonics

A fault signal has been shown in Fig.2. This signal has been generated by combining pure sinusoidal wave form with harmonics and with swell as per the data available in Table-1.

Table 1 Categories of Power System Faults						
Category	Duration	Voltage				
Harmonics	Present in the entire signal	0.0035– 0.087 pu				
Short duration variation- Instantaneous Interruption	0.5 – 30 cycles	<0.1 pu				
Short duration variation- Instantaneous Sag(dip)	0.5 – 30 cycles	0.1 - 0.9 pu				
Short duration variation- Instantaneous Swell	0.5 – 30 cycles	1.1 -1.8 pu				
Short duration variation- Momentary Interruption	30 cycles – 3s	<0.1 pu				
Short duration variation- Momentary Sag(dip)	30 cycles – 3s	0.1 - 0.9 pu				
Short duration variation- Momentary Swell	30 cycles – 3s	1.1 -1.8 pu				
Short duration variation- Temporary Interruption	3s – 1min	<0.1 pu				
Short duration variation- Temporary Sag(dip)	3s – 1min	0.1 – 0.9 pu				
Short duration variation- Temporary Swell	3s – 1min	1.1 -1.8 pu				

The Approximation and Details for different conditions of the signals are given in Figure 3 – Figure 5.

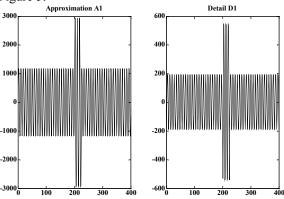


Figure 3 Fifth level decomposition of Instantaneous swell

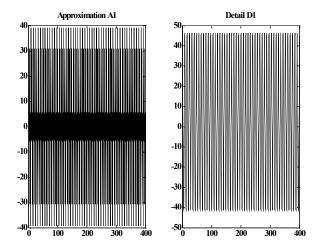


Figure 4 Fifth level decomposition of Harmonics_signal

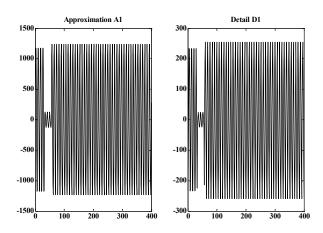


Figure 5 Fifth level decomposition of Instantaneous_interruption_point1_3cycles_no_h armonics

It can be noted from the Figures 3-5 that the wavelet decomposition really helps in identifying the presence of disturbances. In order to make it more clear features are extracted from the approximation and details.

The features are obtained from the Approximation and Details of the 5th level by using the following equations

$$V1 = \frac{1}{d} \sum (IF)$$
 (1)

Where d = Samples in a frame and V1 = Mean value of Instantaneous Frequency

$$V2 = \frac{1}{d} \sum (IF - V1) \tag{2}$$

Where V2=Standard Deviation of Instantaneous Frequency

Where V2=Standard Deviation of Instantaneous Frequency

$$V3 = Maximum(IF)$$
 (3)

$$V4 = Minimum(IF)$$
 (4)

$$V5 = norm(IF)^2$$
 (5)

Where V5 = Energy value of frequency

4. Radial basis function

Radial basis function is a supervised neural network. The network has an input layer, hidden layer (RBF layer) and output layer. The 10 features obtained are used as inputs for the network and the target values for training each fault is based on the values given in Table 2.

Training RBF is done as follows,

- 1. Distance between pattern and centers are found
- An RBF matrix whose size will be (np X cp). , where np= number of pattern (50 signals in each fault X number of faults) used for training and cp is number of centers which is equal to 10.
- 3. Final weights are calculated.
- 4. During testing the performance of the RBF network, RBF values are formed from the features obtained from a signal and processed with the final weights obtained during training. Based on the result

obtained, the electrical fault signal is classified.

5. Experiment and Results

Matlab 2009 has been used to generate 300 signal patterns for each electrical fault. The equation used for generating an electrical fault signal is as follows:

$$y = A \sin(2 \pi f t + \varphi)$$
 (6)

Where $f = Frequency of the signal, A = Amplitude of the signal and <math>\phi = Phase angle$.

Only one wave form is created when considering pure signal,. Different harmonics with different amplitudes are generated and summed. The harmonics waveform is added to fault signals. Harmonics signals are generated with f_h (1st to 25th harmonics) and voltage of the harmonics, v_h (0.004 pu to 0.09pu). Similarly for each fault, swell, sag, interruption under instantaneous, momentary and transient, 300 samples are created with varied time duration, varying amplitudes with 50 Hz as constant frequency given in Table 2. Hence, a total of (10 faults * 300) + 1 pure signals are generated using equation (6). At random, 20 signals are considered from each fault for training RBF. They are given in Table 2.

Table 2 Faults Labeling							
	Number of patterns used for testing	Number of patterns used for training	Targe t label ing				
Pure sine wave	1	1	0.01				
Instantaneous Interruption	250	50	0.02				
Momentary Interruption	250	50	0.03				
Temporary Interruption	250	50	0.04				
Temporary sag	250	50	0.05				
Instantaneous sag	250	50	0.06				
Momentary sag	250	50	0.07				
Temporary swell	250	50	0.08				
Momentary swell	250	50	0.09				
Instantaneous swell	250	50	0.1				
Harmonics	250	50	0.11				

A counter is used to make a note in how many frames have the same values of frequency and the voltage occurs continuously If duration of existence of fault is in multiples of 8, 16, 24, 32, 40, 48 cycles or more than one second then. Based upon the number of similar values in the successive frames, a

fault is classified. In a signal of 1 sec, minimum 2 faults are introduced along with harmonics.

6. Conclusion:

In this work, wavelet (db1) has been used to obtain Approximation and Details at the 5th level decomposition. As the initial sample size is 12800 in the sampled signal, in the 5th level decomposition, 400 samples for Approximation and 400 samples for Details are obtained. The features mean, std, norm, minimum and maximum of the Approximation and Details are obtained. The features are used as training and testing data for the RBF network. The percentage of electrical fault identification in given in Table 3

Table 3 Electrical fault identification							
Faults	Number of patterns identified	Number of patterns used for testing	% identification and classification				
Pure sine wave	1	1	100				
Instantaneous Interruption	235	250	94				
Momentary Interruption	238	250	95.2				
Temporary Interruption	247	250	98.8				
Temporary sag	241	250	96.4				
Instantaneous sag	238	250	95.2				
Momentary sag	247	250	98.8				
Temporary swell	238	250	95.2				
Momentary swell	247	250	98.8				
Instantaneous swell	247	250	98.8				
Harmonics	241	250	96.4				

References

- [1] P.K. Dash, R.K. Jena, M.M.A. Salama, Power quality monitoring using an integrated Fourier linear combiner and fuzzy expert system, *Electrical Power and Energy Systems*, *21*, 1999, 497–506.
- [2] S. Suja, Jovitha Jerome, Pattern recognition of power signal disturbances using S Transform and TT Transform, *Electrical Power and Energy Systems*, *32*, 2010, 37-53.
- [3] P.K. Dash, S.K. Panda, A.C. Liew, B. Mishra, R.K. Jena, A new approach to monitoring electric power quality, *Electric Power Systems Research*, 46, 1998. 11–20.
- [4] Mário Oleskovicz, Denis V. Coury, Odilon Delmont Felho, Wesley F. Usida, Adriano A.F.M. Carneiro, Leandro, R.S. Pires, Power quality analysis applying a hybrid methodology with wavelet transforms and neural networks, *Electrical Power and Energy Systems*, 31, 2009, 206–212.
- [5] S.A. Soliman, M.E. El-Hawary, Measurement of power systems voltage and flicker levels for power quality analysis: a static LAV state estimation based algorithm, *Electrical Power and Energy Systems*, 22, 2000, 447–450.
- [6] S.A. Soliman, A.H. Mantaway, M.E. El-Hawary, Simulated annealing optimization algorithm for power systems quality analysis, *Electrical Power and Energy Systems*, 26, 2004, 31–36.
- [7] M. Aiello, A. Cataliotti, S. Nuccio, A Chirp-Z transform-based synchronizer for power system measurements, *IEEE Trans. Instrum. Meas*, *54*, 2005, 1025–1032.
- [8] S. Santoso, E. J. Powers, W. M. Grady, and A. C. Parsons, "Power quality disturbance waveform recognition using wavelet-based neural classifier—part 1: theoretical foundation," *IEEE Trans. Power Delivery*, vol.15, pp. 222–228, Jan. 2000.
- [9] , "Power quality disturbance waveform recognition using waveletbased neural classifier—part 2: application," *IEEE Trans. Power Delivery*, vol. 15, pp. 229–235, Jan. 2000.
- [10] J. Chung, E. J. Powers, J. Lamoree, and S. C. Bhatt, "Power disturbance classifier using a rule-based method and wavelet packet-based hidden Markov model," *IEEE Trans. Power Delivery*, vol. 17, pp. 233–241, Jan. 2002.
- [11]G. Zheng, D. M'x' Shi, Liu,1. Yao, Miao, Z.M, Power quality disturbance classification based on rule-based and wavelet-multi-resolution decomposition, *Proceedings of 2002 International Conference on Machine Learning and Cybernetics*, 4, 2002, 2137-2141.
- [12] F. Mo and W. Kinsner, "Wavelet modeling of transients in power systems," in *Proc. Conf.*

- Communications, Power and Computing Proceedings, Winnipeg, MB, Canada, May 22–23, 1997, pp. 132–137.
- [13] S. Santoso, E. J. Powers, W. M. Grady, and P. Hofmann, "Power quality assessment via wavelet transform analysis," *IEEE Trans. Power Delivery*, vol. 11, pp. 924–930, Apr. 1996.
- [14] S. Santoso, W. M. Grady, E. J. Powers, J. Lamoree, and S. C. Bhatt, "Characterization of disturbance power quality events with Fourier and wavelet transforms," *IEEE Trans. Power Delivery*, vol. 15, pp. 247–254, Jan. 2000.
- [15] A. M. Gaouda, M. M. A. Salama, M. R. Sultan, and A. Y. Chikhani, "Power quality detection and classification using wavelet-multiresolution signal decomposition," *IEEE Trans. Power Delivery*, vol. 14, pp. 1469–1476, Oct. 1999.
- [16]Suriya Kaewarsa. Kitti Attakitmongcol. Thanatchai Kulworawanichpong, Recognition of power quality events by using multiwavelet-based neural networks, *Electrical Power and Energy Systems*, 30, 2008, 254–260.
- [17] B. Perunicic, M. Mallini, Z. Wang, and Y. Liu, "Power quality disturbance detection and classification using wavelets and artifical neural networks," in *Proc. 8th Int. Conf. Harmonics and Quality of Power*, vol. 1,Oct. 14–16, 1998, pp. 77–82.
- [18]T. Jayasree, D. Devaraj, R. Sukanesh, Power quality disturbance classification using Hilbert transform and RBF networks, *Neurocomputing*, 73, 2010, 1451-1456.
- [19] A. Elmitwally, S. Farghal, M. Kandil, S. Abdelkader, and M. Elkateb, "Proposed wavelet-neurofuzzy combined system for power quality violations detection and diagnosis," *Proc. Inst. Elect. Eng., Gen. Transm. Dist.*, vol. 148, no. 1, pp. 15–20, Jan. 2001.
- [20] L. Angrisani, P. Daponte, M. D' Apuzzo, and A. Testa, "A measurement method based on thewavelet transform for power quality analysis," *IEEE Trans. Power Delivery*, vol. 13, pp. 990–998, Oct. 1998.

GA-ANN based Dominant Gene Prediction in Microarray Dataset

Manaswini Pradhan

Lecturer, P.G. Department of Information and Communication Technology, Fakir Mohan University, Orissa, India

Dr. B. Mittra

Reader, School of Biotechnology, Fakir Mohan University, Orissa, India

Abstract-Genome Analysis of a human being permits useful insight into the ancestry of that person and also facilitates the determination of weaknesses and susceptibilities of that person towards inherited diseases. The amount of accumulated genome data is increasing at a tremendous rate with the rapid development of genome sequencing technologies and gene prediction is one of the most challenging tasks in genome analysis. Many tools have been developed for gene prediction which still remains as an active research area. Gene prediction involves the analysis of the entire genomic data that is accumulated in the database and hence scrutinizing the predicted genes takes too much of time. However, the computational time can be reduced and the process can be made more effective through the selection of dominant genes. In this paper, a novel method is presented to predict the dominant genes of ALL/AML cancer. First, to train an FF-ANN a combinational data of the input dataset is generated and its dimensionality is reduced through Probability Principal Component Analysis (PPCA). Then, the classified database of ALL/AML cancer is given as the training dataset to design the FF-ANN. After the FF-ANN is designed, the genetic algorithm is applied on the test input sequence and the fitness function is computed using the designed FF-ANN. After that, the genetic operations crossover, mutation and selection are carried out. Finally, through analysis, the optimal dominant genes are predicted.

Keywords- gene prediction, Microarray gene expression data, Probabilistic PCA (PPCA), dimensionality reduction, Artificial Neural Network (ANN), Back propagation (BP), dominant gene, genetic algorithm.

I. INTRODUCTION

In the public domain huge quantity of genomic and proteomic data are accessible. The capability to process this information in ways that are helpful to humankind is becoming more and more significant [1]. A fundamental

Dr. Sabyasachi Pattnaik

Reader, P.G. Department of Information and Communication Technology, Fakir Mohan University, Orissa, India.

Dr. Ranjit Kumar Sahu

Assistant Surgeon, Post Doctoral Department of Plastic and Reconstructive Surgery, S.C.B. Medical College, Cuttack, Orissa, India

step in the understanding of a genome is the computational recognition, and in the analysis of newly sequenced genomes it is one of the challenges. Accurate and speedy tools are essential for the analysis of genomic sequences and for interpreting genes [2]. In such circumstances, conventional and modern signal processing techniques plays a vital part in these fields [1]. Genomic signal processing [11] (GSP) is a comparatively novel area in bio-informatics. It deals with the utilization of traditional digital signal processing (DSP) techniques in the representation and analysis of genomic data.

The code for the chemical composition of a particular protein is enclosed in the DNA which is a segment of gene. Genes functions as the pattern for proteins and some extra products, and the main intermediary that translates gene information in the production of genetically encoded molecules is mRNA [4]. Usually sequences of nucleotide symbols, symbolic codons (triplets of nucleotides), or symbolic sequences of amino acids in the corresponding polypeptide chains present in the strands of DNA molecules represent the genomic information. [2]. Gene expression microchip, which is perhaps the most rapidly expanding tool of genome analysis enables simultaneous monitoring of the expression levels of tens of thousands of genes under diverse experimental conditions. An influential tool in the study of collective gene reaction to changes in their environments is presented by gene expression microchip, and it also offers indications about the structures of the involved gene networks [3].

Nowadays, in a solitary experiment by employing microarrays the expression levels of thousands of genes, possibly all genes in an organism can be measured simultaneously [4]. In monitoring genome-wide expression levels of gene microarray technology has become a requisite tool [5]. The evaluation of the gene expression profiles in a

variety of organs which employs microarray technologies disclose separate genes, gene ensembles, and the metabolic ways underlying the structural and functional organization of an organ and its physiological function [6]. By the employment of microarray technology the diagnostic chore can be automated and the precision of the conventional diagnostic techniques can be enhanced. Simultaneous examination of thousands of gene expressions is being facilitated by microarray technology [7].

Efficient representation of cell characterization at the molecular level is possible with microarray technology which simultaneously measures the expression levels of tens of thousands of genes [8]. Gene expression analysis [10] [12] that utilizes microarray technology has a broad variety of latent for discovering the biology of cells and organisms [9]. Accurate prediction and diagnosis of diseases is been assist by the microarray technology. For envisaging the entire gene structure, mainly the precise exon-intron structure of a gene in a eukaryotic genomic DNA sequence gene identification is employed. After sequencing, finding the genes is one of the first and most significant steps in knowing the genome of a species [13]. A field of computational which involved biology is algorithmically distinguishing the stretches of sequence, generally genomicDNA that are biologically functional is known as gene finding. This in particular not only engrosses protein-coding genes but also includes added functional elements for instance RNA genes and regulatory regions [14]. Some of the researches on the gene prediction are [15], [16], [17] and [18].

In this paper, we propose an effective gene prediction technique which predicts the dominant genes. Initially, the classified microarray gene dataset (either Acute Myeloid Leukemia (AML) or Acute Lymphoblastic Leukemia (ALL)) which is of high dimension is reduced through the Probability Principal Component Analysis (PPCA) to generate the training dataset for the neural network. Consequently, through the training data the Feed Forward-ANN is designed and then the genetic algorithm is utilized to predict the dominant genes of ALL/AML cancer. Subsequently the gene which causes either AML or ALL is predicted devoid of analyzing the entire database. The rest of the paper is organized as follows. Section 2 details the genetic algorithm and in Section 3, a brief review of some of the existing works in gene prediction is presented. The proposed effective gene prediction is detailed in Section 4. Section 5 describes the results and discussion. The conclusions are summed up in Section 6.

II. GENETIC ALGORITHM

The heredity and evolution of living organisms are stimulated by computer programs known as Genetic Algorithms [27]. By utilizing GAs an ideal solution is possible even for multi modal objective functions because they are multi-point search methods. Moreover, GA's are applicable to distinct problem in the search space. Hence, GA is not only very simple to use but also a very powerful optimization tool [28]. Strings are present in the search space of GA, each of which represents a candidate solution to the problem and are termed as chromosomes. Fitness value is the objective function value of each chromosome. A set of chromosomes along with their associated fitness is termed as population. The populations which are generated in an iteration of the genetic algorithm are termed as generations [29].

New generations (offspring) are generated by utilize crossover and mutation techniques. Two chromosomes are split by crossover and by taking one split part from each chromosome and combining those two new chromosomes are created. A single bit of a chromosome is changed by mutation. The chromosomes with the best fitness value calculated for a certain fitness criteria are retained while the other chromosomes are removed. The process is repeated until one chromosome has the best fitness value and that chromosome is selected as the solution for the problem [30].

III. REVIEW ON RELATED RESEARCHES

A handful of recent research works available in the literature are briefly reviewed in this section.

A computational technique for patient outcome prediction was introduced by Huiqing Liu *et al.* [19]. Two extreme types of patient samples were utilized for the training phase of this technique:

- 1) short-term survivors who got an inopportune result in a small period and
- 2) long-term survivors who were preserving a positive outcome after a long follow-up time.

These incredible training samples generated a clear platform for identifying suitable genes whose expression was intimately related to the outcome. With the assistance of a support vector machine the selected extreme samples and the important genes were then integrated in order to construct a prediction model. Every validation sample is owed a risk score that falls into one of the special predefined risk groups by employing that prediction model. Several public datasets adapts this technique. In quite a few cases as perceived in their Kaplan–Meier curves, patients in high and low risk groups who are rated by the suggested technique have obviously clear outcome position. They have also established that for enhancing the prediction accuracy, the suggestion of deciding merely extreme patient samples for training is efficient when diverse gene selection techniques are employed.

MiTarget which is a SVM classifier for miRNA target gene prediction was introduced by Kim *et al.* [20]. It employed a radial basis function kernel and was then

categorized by structural, thermodynamic, and position-based features as a similarity measure for SVM features. For the first time, the features were presented and the mechanism of miRNA binding was reproduced. When compared with previous tools the SVM classifier has created high performance with the assistance of biologically pertinent data set that was attained from the literature. The important tasks for human miR-1, miR-124a, and miR-373 was computed by employing Gene Ontology (GO) analysis and the importance of pairing at positions 4, 5, and 6 in the 5' region of a miRNA was explained from a feature selection experiment. A web interface for the program was also presented by them.

Based on the information that a majority of exon sequences have a 3-base periodicity, and intron sequences do not have the sole characteristic, a technique to predict protein coding regions was developed by Changchuan Yin et al. [21]. By employing nucleotide distributions in the three codon positions of the DNA sequences this technique computed the 3-base periodicity and the background noise of the stepwise DNA segments of the target DNA sequences. From the trends of the ratio of the 3-base periodicity to the background noise in the DNA sequences the exon and intron sequences can be recognized. Case studies on genes from diverse organisms illustrated that the proposed technique was an efficient means for exon prediction

On the basis of a two-stage machine learning approach a gene prediction algorithm for metagenomic fragments was proposed by Hoff et al. [22]. Initially, for extracting the features from DNA sequences, linear discriminants were employed for monocodon usage, dicodon usage and translation initiation sites. Secondly, for calculating the chance in such a way that the open reading frame encodes a protein and an artificial neural network combines these characteristics with open reading frame length and fragment GC-content. This probability was employed for categorizing and achieving the gene candidates. By means of extensive training this technique formed fast single fragment predictions with fine quality sensitivity and specificity on artificially fragmented genomic DNA. Additionally, with high consistency this technique can precisely calculate translation initiation sites and distinguish complete genes from incomplete genes. Extensive machine learning techniques were well-suited for predicting the genes in metagenomic DNA fragments. Specially, the association of linear discriminants and neural networks was a very promising one and are believed to be taken into consideration for incorporating into metagenomic analysis pipelines.

Based on the physicochemical features of codons computed from molecular dynamics (MD) simulations an ab initio model for gene prediction in prokaryotic genomes was introduced by Poonam Singhal *et al.* [15]. For every codon

the model requires a statement of three computed quantities, the double-helical trinucleotide base pairing energy, the base pair stacking energy, and a codon propensity index for protein-nucleic acid interactions. Fixing these three parameters, for each codon, eases the computation of the magnitude and direction of a cumulative three-dimensional vector for any length DNA sequence in all the six genomic reading frames. Analysis of 372 genomes containing 350,000 genes has confirmed that the orientations of the gene and non-gene vectors were significantly apart and a apparent difference was made probable between genic and non-genic sequences at a level comparable to or superior than currently accessible knowledge-based models trained on the basis of empirical data, providing a strong evidence for the likelihood of a unique and valuable physicochemical classification of DNA sequences from codons to genomes.

For the genus Aspergillus a program called NetAspGene which is a dedicated, publicly available, splice site prediction was developed by Kai Wang *et al.* [23]. The most widespread mould pathogen that is the gene sequences from Aspergillus fumigatus, were employed to build and test their model. Aspergillus encloses smaller introns when compared with several animals and plants; and hence to cover both the donor and acceptor site information they have applied a larger window size on single local networks for training. NetAspGene was applied to other Aspergilli, including Aspergillus nidulans, Aspergillus oryzae, and Aspergillus niger. Valuation with independent data sets disclosed that NetAspGene executed significantly better splice site prediction than the other available tools.

Bayesian kernel was represented for the Support Vector Machine (SVM) by Alashwal et al. [24] so as to predict protein-protein interactions. By putting together the probability characteristic of the existing experimental protein-protein interactions data, the classifier performances that were amassed from diverse sources could be improved. In addition to that, so as to organize more research on the highly estimated interactions, the biologists are enhanced with the probabilistic outputs that are attained from the Bayesian kernel. The results have illustrated that by employing the Bayesian kernel when compared with the standard SVM kernels, the precision of the classifier has been enhanced. Those results have suggested that by means of Bayesian kernel, the protein-protein interaction could be computed with superior accuracy as when compared to the standard SVM kernels.

IV. PROPOSED DOMINANT GENE PREDICTION USING GENETIC ALGORITHM

Generally, utilization of large gene dataset for disease analysis increases the computation time and degrades the performance of the process. Hence, a technique that requires less computational time to predict dominant genes is essential. Hence, an efficient technique is proposed

to predict the dominant genes of cancer (either AML or ALL) from a microarray gene dataset. The three phases involved in the proposed technique are generation of training dataset, training through neural network and genetic algorithm based dominant gene prediction. Preprocess of dominant gene prediction process is illustrated in Fig. 1 and the feed forward neural network is depicted in Fig. 2.

A. Preprocess for dominant gene prediction

The pre processing steps for predicting dominant genes are explained in the following steps.

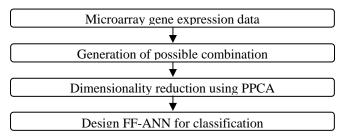


Fig.1 preprocessing steps for dominant gene prediction

1) Generation of training dataset

In this phase, in order to generate the training set for the ANN, it is essential to generate the possible combinations of the gene dataset. The two processes involved in the generation of training dataset are generation of possible combinational data and dimensionality reduction

Possible combinational data are generated by classifying the microarray gene dataset with a lot of combinations within the dataset. This combinational data is generated with the intention of making easier the learning process for dominant genes prediction. Let M_{ij} be the microarray gene dataset, where $0 \le i \le N_s - 1$ and $0 \le j \le N_g - 1$. Here, N_s represents the number of samples and N_g represents the number of genes and the size of M_{ij} is given by $N_s \times N_g$. The number of possible combinational data is calculated as follow,

No of possible combinations =
$$\frac{(N_s \times N_g)!}{((N_s \times N_g) - k)!k!}$$
 (1)

The combinational data $M_{c_{ij}}$ has a high dimension of $N_s \times N_g$ which has to be reduced so as to be utilized in further processing.

2) Dimensionality reduction by PPCA

The dimension of the $M_{c_{ij}}$ must be reduced for the upcoming processes. The dimensionality reduction is done utilizing the probabilistic Principal Component Analysis (PCA) and the high dimensional $M_{c_{ij}}$ was converted to low dimension. The dimensionality reduced data is utilized as the training dataset for the neural network. We reduce the dimensionality using PPCA, which is a PCA that has a probabilistic model for the data. The PPCA algorithm which was composed by Tipping and Bishop [25] utilizes a rightly formed probability distribution of the higher dimensional data and calculates a low dimensional representation.

The instinctive attraction of the probabilistic representation is because of the fact that the definition of the probabilistic measure allows comparison with other probabilistic techniques, at the same time making statistical testing easier and permitting the utilization of Bayesian methods. By making use of PPCA as a generic Gaussian density model dimensionality reduction can be achieved. Efficient computation of the maximum-likelihood estimates for the parameters connected with the covariance matrix from the data principal components is facilitated through dimensionality reduction. The combinational data $M_{c_{ii}}$ of dimension $N_s \times N_g$ is reduced through the PPCA to $\hat{M}_{c_{ij}}$ of dimension $N_s^{"}\times N_g^{"}.$ In addition to dimensionality reduction, the PPCA finds more practical advantages such as finding missing data, classification and novelty detection [26]. Thus training dataset $\hat{M}_{c_{ii}}$ for the ANN is generated with reduced dimension N_s " $\times N_\varrho$ ".

B. Training phase: Training through Feed Forward ANN

The proposed technique incorporates a multilayer feed forward ANN with back propagation for predicting the dominant genes of the AML/ALL cancer. A feed-forward network maps a set of input values to a set of output values and can be thought of as the graphical representation of a parametric function. The dimensionality reduced microarray gene dataset is utilized for training the feed forward Neutral network with back propagation.

The single network N is trained in our proposed approach; the network is for receiving the dimensionality reduced gene dataset, and outputs the gene value whether it is ALL/AML. Hence, the network is configured with $N_{\rm g}$ input units and hidden and an output unit.

Step 1: As the first step, set the input weights of every neuron, apart from the neurons in the input layer.

Step 2: A neural network with N_g input layers, a N_g hidden layers and an output layer are designed. In this neural network, N_s (dimensionality reduced) input neurons and a bias neuron, N_g hidden neurons and a bias neuron and an output neuron y_i are presented.

Step 3: The designed NN is weighted and biased. The developed NN is shown in the Fig.2.

Step4: The basis function and the activation function which is chosen for the designed NN are shown below.

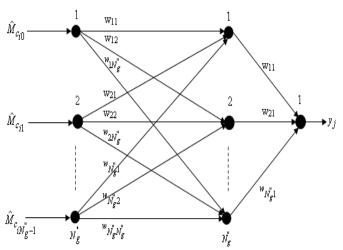


Fig 2. *n* Inputs one output Neural Network to train the gene

$$Y_i = \alpha + \sum_{j=0}^{N_g^{"}-1} w_{ij} \hat{M}_{c_{ij}}, \quad 0 \le i \le N_s^{"}-1$$
 (2)

$$g(y) = \frac{1}{1 + e^{-y}} \tag{3}$$

$$g(y) = y \tag{4}$$

Eq.(2) is the basis function for the input layer, where \hat{M}_c is the dimensionality reduced microarray gene data, w_{ij} is the weight of the neuron and α is the bias. The sigmoid function for the hidden layer is given in Eq.(3) and the activation function for the output layer is given in Eq.(4). The basis function given in Eq. (1) is commonly used in all

the remaining layers (hidden and output layer, but with the number of hidden and output neurons, respectively). The output of the ANN is determined by giving it \hat{M}_c as the input.

Step 5: The learning error is determined for the NN as follows

$$E = \frac{1}{N_s''} \sum_{b=0}^{N_s''-1} D - Y_b$$
 (5)

Here, E is the error in the FF-ANN, D is the desired output and Y_b is the actual output.

1) Minimization of Error by BP algorithm

The steps involved in training BP algorithm based NN is given below.

- a) Randomly generated weights in the interval [0,1] are assigned to the neurons of the hidden layer and the output layer. But all neurons of the input layer have a constant weight of unity.
- b) In order to determine the BP error using Eq. (5), the training gene data sequence is given to the NN. Eq. (2), Eq. (3) and Eq. (4) show the basis function and transfer function.
- c) The weights of all the neurons are adjusted when the BP error is determined as follows,

$$W_{ij} = W_{ij} + \Delta W_{ij} \tag{6}$$

The change in weight Δw_{ij} given in Eq. (6) can be determined as $\Delta w_{ij} = \gamma.y_{ij}$. E, where E is the BP error and γ is the learning rate, normally it ranges from 0.2 to 0.5.

d) After adjusting the weights, steps (b) and (c) are repeated until the BP error gets minimized. Normally, it is repeated till the criterion, E < 0.1 is satisfied.

When the error gets minimized to a minimum value it is construed that the designed ANN is well trained for its further testing phase and the BP algorithm is terminated. Thus, the neural network is trained by using the samples. Then to determine the dominant genes of the ALL/AML cancer the genetic algorithm is applied.

C. Testing phase: Genetic Algorithm based dominant gene prediction of AML/ALL cancer

In the training phase, by means of the training dataset the FF-ANN is designed and the well trained network is utilized for predicting the dominant genes in an efficient manner. The genetic algorithm is applied on the classified test sequence and then this test sequence is evaluated and the dominant genes are predicted. In this GA based dominant gene prediction, initially, the random chromosomes are generated. The random chromosomes are the indices of the test sequence which are classified as ALL/AML. The genes are generated without any repetition within the chromosome. After generating the chromosomes, the fitness is calculated by providing the genes of the chromosome which are the indices as input to the designed FF-ANN. Then, by subjecting the chromosomes to the genetic operations, crossover and mutation, newly generated chromosomes are obtained. Then the fitness is determined for the newly generated chromosomes. The generated new chromosomes are given as input to the designed FF-ANN. The optimal chromosomes are obtained by analyzing the threshold value. The process is repeated until optimal gene values are obtained. The process of genetic algorithm to predict the dominant gene is depicted in fig.3

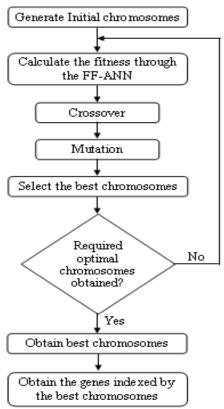


Fig 3. Proposed genetic algorithm for dominant gene prediction

1) Generation of chromosomes

Initially generate N_p number of random chromosomes and the number of genes in each chromosome

relies on N_g i.e. number genes in the training dataset. As discussed earlier, the generated genes are the indices of the test input sequence.

$$D^{(k)} = \left\{ D_0^{(k)}, D_2^{(k)}, D_3^{(k)}, \dots, D_{n-1}^{(k)} \right\}$$

$$0 \le k \le N_p - 1 \quad 0 \le l \le n - 1$$
(7)

n- Number of genes in the training dataset.

In eq.7, $D_l^{(k)}$ represents the l^{th} gene of the k^{th} chromosome. These genes are generated without any repetition within the chromosomes. Once the N_p chromosomes are generated then the fitness function is applied on the generated chromosomes

2) Fitness Function

The fitness of the generated chromosomes is evaluated using the fitness function by giving the chromosomes as input to the designed FF-ANN.

$$\mu_{net} = \frac{\sum_{k=0}^{N_p - 1} N_{out}}{|k|}$$

$$(8)$$

$$N_{fit} = \frac{1}{(1 - \mu_{net})^{c}}$$

$$c = 0 \text{ if test sequence is ALL}$$

$$c = 1 \text{ if test sequence is AML}$$
(9)

In Eq. (8), N_{out} is the network output obtained from the FF-ANN for the k^{th} chromosome and N_{fit} in Eq. (9) is the fitness value of the initially generated chromosomes.

3) Crossover and Mutation

The two point crossover is chosen with the crossover rate of C_R amid diverse kinds of crossovers. Using eq. (10) and (11) two points are selected on the parent chromosomes in the two point crossover. The genes that are present in between the two points cr_1 and cr_2 are exchanged among the parent chromosomes, hence N_p

children chromosomes are attained. The crossover points cr_1 and cr_2 are determined as follows

$$cr_1 = \frac{|l|}{3} - 2 \tag{10}$$

$$cr_2 = \frac{|I|}{2} + 2 \tag{11}$$

The children chromosomes are acquired now and their corresponding gene values are store discretely and their corresponding indices from the $D_l^{(k)}$ are stored in D_{newl}^{-k} . Subsequently mutation is executed by employing Eq. (9) on the chromosomes that are obtained after crossover. Then, by reinstating N_m number of genes from every chromosome with new genes, mutation is achieved. The N_m numbers of gene are just genes, which have the least N_{out} (as determined from the Eq. (9)). The arbitrarily generated genes are the replaced genes devoid of any recurrence within the chromosome. Then, the selected chromosomes for crossover operation, and the chromosomes which are obtained from mutation are combined, hence the population pool is filled up with the N_p chromosomes. Then, until a maximum iteration of I_{max} is reached this process is repeated iteratively.

4) Selection of optimal solution

The best chromosomes are selected from the group of chromosomes that is obtained after the process is repeated $I_{\rm max}$ times. Here, the best chromosomes are the chromosomes which have minimum fitness for both ALL/AML which may depend upon the c value. The obtained best chromosomes are used to retrieve the corresponding gene values from the test sequence. The gene values of the ALL/AML cancer represented by the indices, which are obtained from the genes of the best chromosomes, are the dominant genes of the ALL/AML and they are retrieved in an effective manner.

V. IMPLEMENTATION RESULTS AND DISCUSSION

The proposed dominant gene prediction technique is implemented in the MATLAB platform (Version 7.10) and it is evaluated using the classified microarray gene expression data of human acute leukemias. The standard leukemia dataset for training and testing is obtained from [26]. The training leukemia dataset is of dimension $N_g=7192$ and $N_s=38$. This dimension of the dataset is too high to train the FF-ANN and hence its dimension is reduced using PPCA and then the training dataset of

dimension $N_g = 30$ and $N_s = 38$ is obtained. This training dataset is utilized to design the FF-ANN and then the test input sequence is tested through the genetic algorithm. The selected double point crossover points are $cr_1 = 8$ and $cr_2 = 22$ with a crossover rate $C_R = 0.5$ and for mutation $N_{\scriptscriptstyle m}=5$. After the completion of the crossover and mutation operations, based on the conditions given in section 4, the optimal chromosomes were obtained. These optimal chromosomes are the indices of the ALL cancer test sequence. This process is repeated until it reaches the maximum iteration $I_{\text{max}} = 20$. The training of FF-ANN is implemented using the Neural Network Toolbox in MATLAB. Fig 4 shows the Regression of the designed FF-ANN and the Fig 5 shows the performance of the designed FF-ANN. Fig 6 depicts the performance of the ALL test sequence during the testing process and the Fig 7 depicts the performance of the AML test sequence during the testing process.

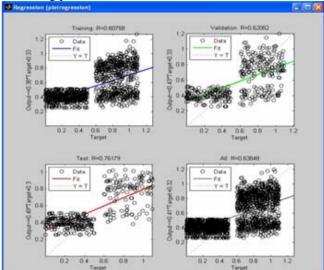


Figure 4: Regression output of the designed FF-ANN

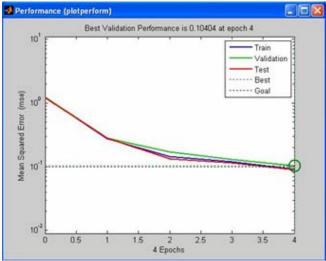


Figure 5: Performance of BP in training the designed FF-ANN

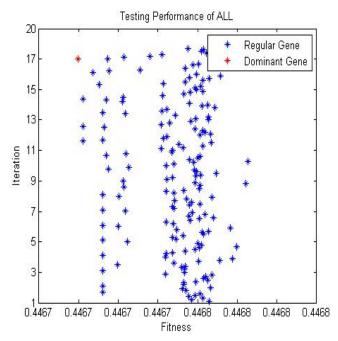


Figure 6: The performance of ALL during the testing process

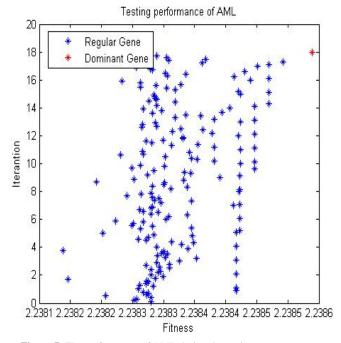


Figure 7: The performance of AML during the testing process

Once the training process of the FF-ANN is completed, the input sequence either ALL or AML is tested through the genetic algorithm and then the dominant gene of either ALL or AML has been obtained. In Fig 6, the performance of the ALL input sequence has been tested and the obtained dominant gene based on some criteria (mentioned in the section 4) is depicted differently from the regular genes. In Fig 7, the performance of the AML input

sequence has been tested and the obtained dominant gene based on some criteria (mentioned in the section 4) is depicted differently from the regular genes. The table 1 demonstrated the dominant genes of the ALL and AML below

ALL AML					
Indices	Dominant Genes	Fitness by FF- ANN	Indices	Dominant Genes	Fitness by FF- ANN
6041	1284		3196	-162	
6378	-231		647	119	
3845	-11		1024	12450	
5764	36		2269	757	
3267	390		4108	177	
518	1396	0.4467	1036	910	2.2381
6485	62		1077	1361	
3756	-482		4763	3381	
3812	251		1905	118	
4122	-16		3790	148	

Table 1. The indices of dominant genes, dominant genes and their fitness

VI. CONCLUSION

In this paper, an effective genetic algorithm based method to predict the dominant genes in the ALL/AML dataset was discussed. The proposed technique, instead of analyzing the entire database, analyzed only the dominant genes and hence it has provided the optimal results. The FF-ANN was designed by means of training samples to assess the test sequence in the proposed genetic algorithm. Then, the fitness of the test sequence samples was evaluated through the designed FF-ANN. After that, the test input sequence was evaluated and the dominant genes were predicted through the genetic algorithm. The obtained fitness of the ALL dominant genes through the FF-ANN is 0.4467 and for AML dominant genes is 2.2381. Table 1 demonstrated the dominant genes of the ALL and the AML.

REFERENCES

- [1] Vaidyanathan and Byung-Jun Yoon, "The role of signal processing concepts in genomics and proteomics", Journal of the Franklin Institute, Vol.341, No.2, pp.111-135, March 2004
- [2] Anibal Rodriguez Fuentes, Juan V. Lorenzo Ginori and Ricardo Grau Abalo, "A New Predictor of Coding Regions in Genomic Sequences using a Combination of Different Approaches", International Journal of Biological and Life Sciences, Vol. 3, No.2, pp.106-110, 2007
- [3] Ying Xu, Victor Olman and Dong Xu, "Minimum Spanning Trees for Gene Expression Data Clustering", Genome Informatics, Vol. 12, pp.24–33, 2001
- [4] Anandhavalli Gauthaman, "Analysis of DNA Microarray Data using Association Rules: A Selective Study", World Academy of Science, Engineering and Technology, Vol.42, pp.12-16, 2008
- [5] Chintanu K. Sarmah, Sandhya Samarasinghe, Don Kulasiri and Daniel Catchpoole, "A Simple Affymetrix Ratio-transformation Method Yields Comparable Expression Level Quantifications with cDNA Data", World Academy of Science, Engineering and Technology, Vol. 61, pp.78-83, 2010
- [6] Khlopova, Glazko and Glazko, "Differentiation of Gene Expression Profiles Data for Liver and Kidney of Pigs", World Academy of Science, Engineering and Technology, Vol. 55, pp.267-270, 2009

- [7] Ahmad m. Sarhan, "cancer classification based on microarray gene expression data using DCT and ANN", Journal of Theoretical and Applied Information Technology, Vol.6, No.2, pp.207-216, 2009
- [8] Huilin Xiong, Ya Zhang and Xue-Wen Chen, "Data-Dependent Kernel Machines for Microarray Data Classification", IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), Vol.4, No.4, pp.583-595, October 2007
- [9] Javier Herrero, Juan M. Vaquerizas, Fatima Al-Shahrour, Lucia Conde, Alvaro Mateos, Javier Santoyo Ramon Diaz-Uriarte and Joaquin Dopazo, "New challenges in gene expression data analysis and the extended GEPAS", Nucleic Acids Research, Vol. 32, pp.485–491, 2004
- [10] Sveta Kabanova, Petra Kleinbongard, Jens Volkmer, Birgit Andrée, Malte Kelm and Thomas W. Jax, "Gene expression analysis of human red blood cells", International Journal of Medical Sciences, Vol.6, No.4, pp.156-159, 2009
- [11] Anastassiou, "Genomic Signal Processing," IEEE Signal Processing Magazine, Vol. 18, PP. 8-20, 2001
- [12] Chen-Hsin Chen, Henry Horng-Shing Lu, Chen-Tuo Liao, Chun-houh Chen, Ueng-Cheng Yang and Yun-Shien Lee, "Gene Expression Analysis Refining System (GEARS) via Statistical Approach: A Preliminary Report", Genome Informatics, Vol.14, pp.316-317, 2003.
- [13] Richard A. George, Jason Y. Liu, Lina L. Feng, Robert J. Bryson-Richardson, Diane Fatkin and Merridee A. Wouters, "Analysis of protein sequence and interaction data for candidate disease gene prediction", Nucleic Acids Research, Vol.34, No.19, pp.1-10, 2006
- [14] Skarlas Lambrosa, Ioannidis Panosc and Likothanassis Spiridona, "Coding Potential Prediction in Wolbachia Using Artificial Neural Networks", Silico Biology, Vol.7, pp.105-113, 2007
- [15] Poonam Singhal, Jayaram, Surjit B. Dixit and David L. Beveridge, "Prokaryotic Gene Finding Based on Physicochemical Characteristics of Codons Calculated from Molecular Dynamics Simulations", Biophysical Journal, Vol.94, pp.4173-4183, June 2008
- [16] Freudenberg and Propping, "A similarity-based method for genome-wide prediction of disease-relevant human genes", Bioinformatics, Vol. 18, No.2, pp.110-115, April 2002
- [17] Hongwei Wu, Zhengchang Su, Fenglou Mao, Victor Olman and Ying Xu, "Prediction of functional modules based on comparative genome analysis and Gene Ontology application", Nucleic Acids Research, Vol.33, No.9, pp.2822-2837, 2005
- [18] Mario Stanke and Stephan Waack, "Gene prediction with a hidden Markov model and a new intron submodel ", Bioinformatics Vol. 19, No. 2, pp.215-225, 2003
- [19] Huiqing Liu, Jinyan Li and Limsoon Wong, "Use of extreme patient samples for outcome prediction from gene expression data", Bioinformatics, Vol.21, No.16, pp.3377-3384, 2005
- [20] Sung-Kyu Kim, Jin-Wu Nam, Je-Keun Rhee, Wha-Jin Lee and Byoung-Tak Zhang, "miTarget: microRNA target gene prediction using a support vector machine", BMC Bioinformatics, Vol.7, No.411, pp.1-14, 2006
- [21] Changchuan Yin and Stephen S.T. Yau, "Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence", Journal of Theoretical Biology, Vol.247, pp.687-694, 2007
- [22] Katharina J Hoff, Maike Tech, Thomas Lingner, Rolf Daniel, Burkhard Morgenstern and Peter Meinicke, "Gene prediction in metagenomic fragments: A large scale machine learning approach", BMC Bioinformatics, Vol. 9, No.217, pp.1-14, April 2008.
- [23] Kai Wang, David Wayne Ussery and Soren Brunak, "Analysis and prediction of gene splice sites in four Aspergillus genomes", Fungal Genetics and Biology, Vol. 46, pp.14–18, 2009
- [24] Hany Alashwal, Safaai Deris and Razib M. Othman, "A Bayesian Kernel for the Prediction of Protein-Protein Interactions", International Journal of Computational Intelligence, Vol. 5, No.2, pp.119-124, 2009
- [25] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis", Journal of the Royal Statistical Society, Series B, Vol. 21, No. 3, p.p. 611–622, 1999
- [26]ALL/AML datasets from
- http://www.broadinstitute.org/cancer/software/genepattern/datasets/
- [27] Goldberg, "Genetic Algorithms in search, optimization and machine learning" Addison-Wesly, 1989
- [28] Tomoyuki Hiroyasu, "Diesel Engine Design using Multi-Objective Genetic Algorithm", Technical Report, 2004

[29] Ahmed A. A. Radwan, Bahgat A. Abdel Latef, Abdel Mgeid A. Ali and Osman A. Sadek, "Using Genetic Algorithm to Improve Information Retrieval Systems", World Academy of Science and Engineering Technology, Vol.17, No.2,pp.6-13, May 2006

[30] Bhupinder Kaur and Urvashi Mittal, "Optimization of TSP using Genetic Algorithm", Advances in Computational Sciences and Technology, Vol.3, No.2, pp.119-125, 2010

AUTHORS PROFILE



Manaswini Pradhan received the B.E. in Computer Science and Engineering, M.Tech in Computer Science from Utkal University, Orissa, India.She is into teaching field from 1998 to till date. Now, she is working as a Lecturer in P.G. Department of Information and Communication Technology, Fakir Mohan University, Odisha, India. She is currently persuing the Ph.D. degree in the P.G. Department of

Information and communication Technology, Fakir Mohan University, Odisha, India. Her research interest areas are neural networks, soft computing techniques, data mining, bioinformatics and computational biology.



Dr. Sabyasachi Pattnaik has done his B.E in Computer Science, M Tech.from IIT Delhi. He has received his PhD degree in Computer Science in the year 2003, now working as Reader in the Department of Information and Communication Technology, in Fakir Mohan University, Vyasavihar, Balasore, Odisha, India. He has got 15 years of teaching and

research experience in the field of neural networks, soft computing techniques. He has got 22 publications in national & international journals and conferencesAt present he is involved in guiding 6 scholars in the field of neural networks in cluster analysis, bio-informatics, computer vision & stock market applications. He has received the best paper award & gold medal from Odisha Engineering congress in 1992 and institution of Engineers in 2009.



Dr B Mitra, Reader, School of Biotechnology, F.M.University, Odisha, working in the area of Proteomics and Bio-informatics. He has fifteen years of research experiences and produced research papers in many international journals related to molecular biology, immunotechnology, and proteomics.



Dr Ranjit Kumar Sahu, M.B.B.S, M.S. (General Surgery), M. Ch. (Plastic Surgery). Presently working as an Assistant Surgeon in post doctoral department of Plastic and reconstructive surgery, S.C.B. Medical College, Cuttack, Odisha, India. He has five years of research experience in the field of surgery and published many international papers in Plastic Surgery.

Therapeutic Diet Prediction for Integrated Mining of Anemia Human Subjects using Statistical Techniques

Sanjay Choudhary

Department of Mathematics & Computer Science

Govt. Narmada P.G. Mahavidyalaya

Hoshangabad, India

Kamal Wadhwa

Department of Mathematics & Computer Science Govt. Narmada P.G. Mahavidyalaya Hoshangabad, India Abha Wadhwa

Department of Computer Science & Application
Govt Girls P.G. College
Hoshangabad, India

Anjana Mishra

Department of Mathematics & Computer Science
Govt. Narmada P.G. Mahavidyalaya
Hoshangabad, India

Abstract: Chronic disease anemia [1] occurs when blood doesn't have enough hemoglobin. Hemoglobin is a protein in red blood cells that carries oxygen from lungs to the rest of our body. All the body parts need oxygen. Anemia can starve our body of the oxygen it needs to survive. Possible causes of anemia include low vitamin B_{12} or folic acid intake and some chronic illnesses. But the most common cause is not having enough iron in blood which needs to make hemoglobin. This type of anemia is called iron deficiency anemia.

Data Mining is widely used in database communities because of its wide applicability. One major application area of Data Mining is in therapeutic diet prediction. There are several chronic diseases which can be prevented using nutritive food. This paper presents association and correlation between anemia human subject and its prevention through diet nutrients. The role of diet in preventing and controlling iron deficiency is significant. Due to changes in dietary and life style patterns anemia can become catastrophic, so by predicting proper and sufficient diet nutrients for individuals, we can reduce the impact of anemia on human subjects.

Keywords: - Chronic Disease, Anemia, Diet Nutrients, Clinical System, Correlation, Data Mining

I. Introduction

Data Mining is referred to as Knowledge Discovery from databases[5], a process of nontrivial extraction of implicit previously unknown and potentially useful information from databases, has wide application in information management concepts, query processing, decision making, process control, statistical analysis etc. [4], [5] An association rule mining is an important process in data

mining, which determines the correlation between items belonging to transaction database [6], [7].

Chronic disease is a disease[10] that is long lasting for recurrent. Anemia also becomes a chronic disease if it is not cured timely. It is prolonged, do not resolve spontaneously and are rarely cured completely. If anemia results from a diet which is low in iron, iron rich foods or iron pills may be the doctor suggests. Data Mining technology can be utilized for improving the quality of health care of an individual. There is need for lowering the cost of health care facilities along with quality based treatment especially for poor sections of our society.

Health care organization need to lower cost, raise quality and still remain competitive. IT can be used for patient health care. In the IT driven society the role of IT in health care is well established. We can use data mining techniques for analyzing patient data to generate predictions and knowledge which can in turn be used for fast and better clinical decision making. Medical data mining deals with large amount of data, there fore we need data mining techniques, which can explore hidden patterns in the data sets of the medical domain. These patterns can be utilized for clinical diagnosis and prediction.

The available raw medical data are widely distributed, heterogeneous in nature and voluminous, therefore these data need to be collected in an organized form. This collected data can be then integrated to form as a basis for prediction of nutrients chronic diseases. Data mining technology provides a user oriented approach to extract hidden patterns from the data. Data mining can deal with heterogeneous type of data. Medical data are different from data in other databases. Medical data are heterogeneous and contain text and images also, for example MRI, ECG, etc generate huge amounts of heterogeneous medical data. Knowledge discovery from this type of data can greatly benefit mankind by improved diagnostic techniques.

A. Background

Diet and nutrition are important factors in the promotion and maintenance of good health throughout the entire life course. The chronic diseases related to diet and nutrition are anemia, obesity, diabetes, cardiovascular disease, cancer, osteoporosis etc. Anemia can be defined as a reduction in the hemoglobin, hematocrit or red cell number[10]. The sudden, rapid loss of 30% of the total blood volume often results in death. The burden of chronic disease is rapidly increasing world wide .It has been calculated that in 2001, chronic diseases contributed approximately 60% of the 56.5 million total reported deaths in the world and approximately 46% of the global burden of disease.

The chronic disease problem has effected a large proportion of our population. It has been projected that by 2020 chronic diseases will account for almost three quarters of all deaths worldwide are due to heart diseases. Chronic diseases are largely preventable disease. Research has shown that diet can control chronic diseases to large extent[11]. Modern dietary patterns and physical activity patterns are beneficial in prevention of chronic diseases. The chemical composition of food and physiological response to diet can be discovered. Unraveling the interconnection between diet and health through data mining techniques is the basis of this paper identification of disease and diet associated composite molecular biomarkers will facilitate identification of new molecular targets for development of novel therapeutic agents to address diet related chronic diseases.

B. Data Mining And Statistics

Data Mining is designed[2] to learn future prediction. The Data Mining tool checks the statistical significance of the predicted patterns and reports. The difference between Data Mining and statistics is that Data Mining automates the statistical process requiring in several mining tools. Statistical inference is assumption driven in the sense that a hypothesis is formed and tested against data. Data Mining in contrast is discovery driven. That is, the hypothesis is automatically extracted from the given data. The other reason is Data Mining techniques tend to be more robust for real-world data and also used less by expert users.

C. Statistical Correlation

Through correlation[3] we may find that a change in one variable result in change of second variable.

Whenever there exists a relationship between two variable such that a change in one variable results in a positive or negative change in other and also greater change in one variable results in a corresponding greater change in the other, the relationship is called correlation and the two variables are called correlated.

Two variables are called positively correlated if corresponding to an increase (or decrease) in one variable results in an increase (or decrease) in the other.

Two variables are negatively correlated if corresponding to an increase (or decrease) in one variable results in decrease (or increase) in the other.

II. ASSOCIATION AND CORRELATION TO BUILD RELATIONSHIP

In this paper we are dealing with two distinct systems, anemia human subject and diet prediction. Based on the level of hemoglobin of individual, we have made an attempt to associate human subject and iron requirement. It has been found that average range of hemoglobin for males is 14-16 and for females is 12-14. The iron requirement for average range of hemoglobin individual according to RDA (Recommended Dietary Allowances) table is 28mg for males and 30mg for females. The table given below depicts the fact that the iron requirement of females is more than that of males. We have used the correlation statistical technique for finding the relationship between both the systems.

In every case of anemia[9], the cause should be discovered and treated. In clinical practice, nutritional anemia is commonly associated with overall under-nutrition and a balanced diet should be given. Usually, diet alone is not adequate and therapy with specific supplements – particularly iron – is also needed. Supplements of 10mg (179 micromol) iron daily prevent iron deficiency. Association rule mining[5] uses support confidence framework. Diet can cure anemia, this statement can be justified through support confidence measure. We can augment support confidence framework through correlation. The Bayesian correlation technique has been used to find the correlation between required level of iron (in mg) and range of hemoglobin of individuals.

III. EXPERIMENTAL RESULTS

Sufficient safe and varied food supplies not only prevent malnutrition but also reduce the risk of chronic diseases whereas nutritional deficiency increases the risk of common infectious diseases generally in children. This paper has used data mining techniques to investigate factors that contribute significantly in reducing the risk of chronic disease anemia.

There are two methods of analyzing correlation between variables:

- (i) Karl Pearson's Method
- (ii) Scatter Diagram Method

(i) Use of Karl Pearson's method to find coefficient of correlation: Correlation is the relationship between two or more than two variables. We are finding correlation between anemia human subjects(male and female) and predicted diet. Anemia is a chronic disease caused due to several factors. Among those factors inadequate and malnutrition is a major cause of anemia. We have generated the following correlation Table 1 for male anemia subject with mid value of hemoglobin $\operatorname{range}(X_m)$ and iron intake $(Y_m \, \text{mg})$:-

Hb range:	0-6	6-8	8-12	12-14	14-16
Iron intake:	42	36	34	31	28

TABLE 1.

X_{m}	Y_{m}	$x=X_m$ -	$y=Y_m$	x2	y2	x*y
		M_{x}	$-M_{ m y}$			
3	42	-6.6	7.8	43.56	60.84	-51.48
7	36	-2.6	1.8	6.76	3.24	-4.68

10	34	0.4	-0.2	0.16	0.4	-0.8
13	31	3.4	-3.2	11.56	10.24	-10.88
15	28	5.4	-6.2	29.16	38.44	-33.48
48	171			$\Sigma x^2 =$	$\Sigma y^2 =$	$\Sigma x^*y =$
				91.2	113.16	-101.32

$$M_{Xm} = \sum X_m / n = 48/5 = 9.6$$

$$M_{Ym} = \sum Y_m / n = 171/5 = 34.2$$

The coefficient of correlation for male anemia subjects:-

$$r = \sum x^*y/\sqrt{\sum x^2 * \sum y^2}$$

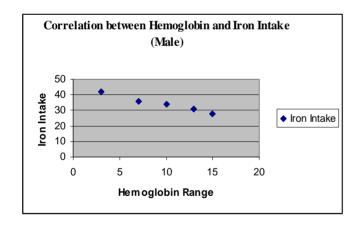
$$=-101.32/\sqrt{(91.2)*(113.16)}$$

$$=-101.32/\sqrt{(10320.192)}$$

$$= -0.9974$$

(ii) *Scatter Diagram Method:* The diagrammatic representation of a bivariate data is known as scatter diagram. For bivariate the values of variables X and Y are plotted (male and female) in the X-Y plane. X axis denote range of hemoglobin and Y axis denotes iron required for individuals.

Figure 1



In this scatter diagram all the points lie on a line and it declines from left to right, this depicts the fact that there is a negative correlation of *high degree* between the variables, i.e. if the hemoglobin count of individual is less than the required iron is more.

We have generated the following correlation Table 1 for female anemia subject with mid value of hemoglobin range(X) and iron intake (Ymg):-

12-14

Hb range: 0-6 6-8 8-12

Iron intake: 42 36 34 30

TABLE 2.

$X_{\rm f}$	$Y_{\rm f}$	$x=X_f$	y= Y _f	x2	y2	x*y
		M_x	-M _y			-
3	45	-5.25	8.5	27.56	72.25	-44.625
7	38	-1.25	1.5	1.56	2.25	-1.875
10	33	1.75	-3.5	3.06	12.25	-6.125
13	30	4.75	-6.5	22.56	42.25	-30.875
33	146			$\Sigma x^2 =$	$\Sigma y^2 =$	$\Sigma x^*y =$
				54.74	129	-83.5

$$M_{Xf} = \sum X_f / n = 33/4 = 8.25$$

$$M_{Yf} = \sum Y_f/n = 146/4 = 36.5$$

The coefficient of correlation for female anemia subjects:-

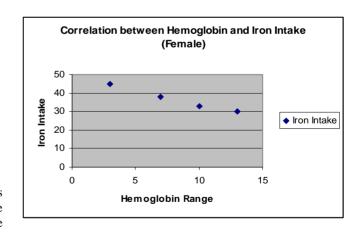
$$r = \sum x*y/\sqrt{\sum x^2*\sum y^2}$$

$$= -83.5/\sqrt{(54.74)*(129)}$$

$$= -83.5/\sqrt{(7061.46)}$$

$$= -0.9937$$

Figure 2



If measure of correlation is between -0.75 and -1, negative then measure of correlation is of high degree.

Both the methods have justified the same fact that there is negative correlation of high degree between hemoglobin range and required iron.

IV. CONCLUSION AND FUTURE SCOPE

Data Mining can be performed by dietitians and hospital administration to prepare diet chart for anemic patients. The goal is

to detect when the Hb percent of patient is very low the diet predicted should contain food high in iron – such as seafood, dried fruits like apricots, prunes, raisins, nuts, beans, green leafy vegetables, whole grains etc..

This research has revealed the fact that anemia human subjects should be recommended food containing high percentage of iron and those having normal range of hemoglobin can take food containing average percentage of iron.

There exist negative correlation[3] between diet high in iron and Hb% of patient. If Hb% of patient is low, diet should be rich in iron and if the reverse condition exist i.e. Hb% of patient is high, diet can contain average percentage of iron.

This paper depicts the correlation between anemia human subject and diet intake of individual. Anemia is common disease effecting large masses of people and if it is not cured timely, it becomes catastrophic. Organisation for social services can use the above result for preventing anemia of particular area. There are several areas which suffer from some kind of deficiencies such as Calcium deficiency is common in Bhopal, Iodine deficiency in Kannur, Purulia Distt., West Bengal, Sickle Cell anemia in tribal population of Maharashtra, Gujrat, Orissa and Tmilnadu etc.. They can use above conclusion to predict diet for anemia affected areas. Data mining deals with large data set, suppose we consider total human population of any area and we can classify the population into male and female data sets. In areas where iron deficiency is common the hemoglobin range of individuals will be below 8%.

This paper has used the correlation concept of statistics to depict and justify the fact that there is high degree of correlation between hemoglobin range of individual and iron intake. By using this fact organizations can plan diet patterns for areas facing the problem of anemia. The entire population of that area can be prescribed iron rich food. Dietitians can also use this fact for preventing this common disease of particular geographical area. There are several remote areas in India which suffer from this common disease. Organisations can directly plan nutrients for those needy people without the need for performing unnecessary calculations. This result will be beneficial for large masses of people. They can be prevented from this chronic disease by giving iron rich food, so the above results are beneficial for our society.

REFERENCES

- [1] The College of Family Physicians of Canada, 2630 Skymark Avenue, Mississauga,
 - ON L4W 5A4
- [2] Jayanthi Ranjan, Applications of Data Mining Techniques in Pharmaceutical Industry, Journal of Theoritical and Applied Information Technology, 2005-2007, 61-67.
- [3] Statistical Methods, H. K. Pathak and Dr. D. C. Agrawal, 258 265, Correlation and Regression.
- [4] Yeong-Chyi Lee A, Tzung-Pei Hong, Tien Chin Wang, "Multi-level fuzzy mining with multiple minimum supports." Journal of Elsevier Expert Systems with Applications, vol.34, pp.459-468, 2008.
- [5] J. Han, M. Kamber, "Data Mining: Concepts and Techniques." The Morgan Kaufmann Series, 261-265, 2001.

- [6] I. Ha, Y. Cai, and N. Cercone, "Data-driven of Quantitative Rules in Relational Databases." IEEE Tram. Knowledge and Data Eng., vol.5, pp. 29-40, 1993.
- [7] R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases." In Proceeding ACM SIGMOD Conference, pp. 207-216, 1993.
- [8] Abdullah H. al-Assaf, "Anemia and Iron Intake of Adult Saudis in Riyadh City-Saudi Arabia." Pakistan Journal of Nutition 6 (4): 355-358, 2007, issn 1680-5194.
- [9] Antia F. P. and Abraham Philip, "Clinical Dietitics and Nutrition", 329-330, 1997.
- [10] Corinne H. Robinson, Marilyn R. Lawler "Normal and Therapeutic Nutrition" 511-519.
- [11] Report of a Joint WHO/FAO Expert Consultation, "DIET NUTRITION AND THE PREVENTION OF CHRONIC DISEASES", WHO Technical Report Series, 916, 4-5.
- [12] Swaminathan M. "Food & Nutrition", 66-74, vol. II, Applied Aspects, 2003.
- [13] Abidi, S.S.R. (2001) Knowledge management in healthcare: towards 'knowledge-driven' decision- support services. *International Journal of Medical Informatics* 63, 5-18.
- [14] B.Shri Laxmi, Food Science.
- [15] Cios, K.J., & Moore, G.W. (2000) Medical Data Mining and knowledge Discovery: An Overview. In Cios K. J. , *Medical Data Mining and knowledge Discovery*. Heidelberg: Physica-Verlag.
- [16] J Am Coll Cardiol. Aug, 28 (2), 515-521. Data Science Journal, Volume 5,19 October 2006.
- [17] J.S. Garrow, Human Nutrition & Dietetics.

Improve The Test Case Design of Object Oriented Software by Refactoring

DIVYA PRAKASH SHRIVASTAVA

Department of Computer Science Al Jabal Al Garbi University Zawya, LIBYA

conceived to deal with source code changes.

Abstract—This Refactoring is the process of changing a software system aimed at organizing the design of source code, making the system easier to change and less error-prone, while preserving observable behavior. This concept has become popular in Agile software methodologies, such as extreme Programming (XP), which maintains source code as the only relevant software artifact. Although refactoring was originally

Two key aspects of eXtreme Programming (XP) are unit testing and merciless refactoring. We found that refactoring test code is different from refactoring production code in two ways: (1) there is a distinct set of bad smells involved, and (2) improving test code involves additional test code refactorings, we describe a set of code smells indicating trouble in test code and a collection of test code refactorings explaining how to overcome some of these problems through a simple program modification.

Keywords- Test Smell, Test Case, Refactoring, Unit Testing, Object Oriented, TDD.

I. INTRODUCTION (HEADING 1)

Computer software is an engine of growth of socialeconomy development which requires new techniques and strategies. The demand for quality in software applications has grown. Hence testing becomes one of the essential components of software development which is the indicator of quality [4].

"Testing proves the presence, not the absence of bugs" -- E.W.Dijkstra

The unit test provides the lowest level of testing during software development, where the individual units of software are tested in isolation from other parts of program/software system. Automated Testing is the other program that runs the program being tested, feeding it with proper input, and thus checking the output against the expected. Once the test case is written, no human intervene is needed thus the test case does all and indicate[12].

Adequate testing of software trials prevent this tragedies to occur. Adequate testing however, can be difficult if the software is extremely large and complex. This is because the amount of time and efforts required to execute a large set of test cases or regression test cases be significant [3]. Therefore,

R.C.JAIN

Department of Computer Application Samrat Ashoka Technological Institute Vidisha. INDIA

the more testing can be done with accuracy of test cases which assist in corresponding rise in program transformation.

Amongst different types of program transformation, behavior-preserving source-to-source transformations are known as refactorings [2]. Refactoring is the process of changing a software system in such a way that it does not alter the external behavior of the code yet improves its internal structure [8].

The refactoring concept was primarily assigned to source code changes. The refactoring of test case may bring additional benefits to software quality and productivity, vis-avis cheaper detection of design flaws and easy exploration of alternative design decisions. Consequently, The term code refactoring and test case refactoring can be made distinct. Thus, one of the main reasons for wide acceptance of refactoring as a design improvement technique and its subsequent adoption by Agile software methodologies, in particular eXtreme Programming (XP) [1]. The XP encourages the development teams to skip comprehensive initial architecture or design stages, guiding them its implementation activities according to user requirements and thus promoting successive code refactorings when inconsistencies are detected.

II. TEST-DRIVEN DEVELOPMENT

Test Driven Development (TDD) is the core part of the Agile code development approach derived from eXtreme Programming (XP) and the principles of the Agile manifesto. It provides to guarantee testability to reach an extremely high test coverage, to enhance developer confidence, for highly cohesive and loosely coupled systems, to allow larger teams of programmers to work on the same code base, as the code can be checked more often. It also encourages the explicitness about the scope of implementation. Equally it helps separating the logical and physical design, and thus to simplify the design, when only the code needed.

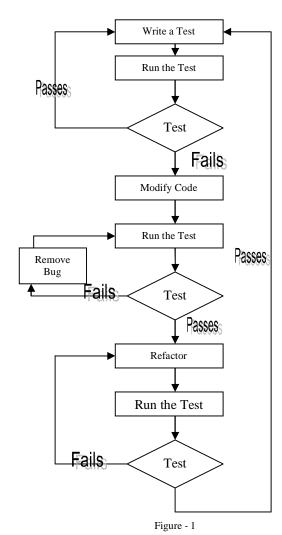
The TDD is not a testing technique, rather a development and design technique in which tests are written prior to the production code. The tests are added its gradually during its implementation and when the test is passed, the code is refactored accordingly to improve the efficacy of

internal structure of the code. The incremental cycle is repeated until all functionality is implemented to final.

The TDD cycle consists of six fundamental steps:

- 1. Write a test for a piece of functionality,
- 2. Run all tests to see the new test to fail,
- 3. Write corresponding code that passes these tests,
- 4. Run the test to see all pass,
- 5. Refactor the code and
- 6. Run all tests to see the refactoring did not change the external behavior.

The first step involves simply writing a piece of code to ensure the tests of desired functionality. The second is required to validate the correctness of test, i.e. the test must not pass at this point, because the behavior under implementation must not exist as yet. Nonetheless, if the test passes over, means the test is either not testing correct behavior or the TDD principles have not been strictly followed. The third step is the writing of the code.



However, it should be kept in mind to only write as little code as possible to enable to pass the test. Next, step is to see that the change has not introduced any of the problems somewhere else in the system. Once all these tests are passed, then the internal structure of the code should be improved by refactoring. The afore mentioned cycle is presented in Fig 1.

III. REFACTORING

The Program restructuring is a technique for rewriting software may be useful either for legacy software as well as for the production of new systems[6, 8,11]. If the internal structure is changed, although the behavior (what the program is supposed to do) is maintained. Restructuring re-organizes the logical structure of source code in order to improve specific attributes [8] or to make it less error-prone when future changes are introduced [11].

Behavior preserving program changes are known as refactorings which was introduced by Opdyke [10]. Yet it's gaining importance by Fowler's work [5] and eXtreme Programming (XP) [1], an Agile software development in context of object-oriented development. In this context, a refactoring is usually composed of a set of small and atomic refactorings, after which the largest source code is better than the original with respect to particular quality attributes, such as readability and modularity.

Thus, refactoring can be viewed as a technique for software evolution through-out software development and maintenance. Software evolution can be classified into the following types [7]:

- Corrective evolution: correction of errors:
- Adaptive evolution: modifications to accommodate requirement changes;
- Perfective evolution: modifications to enhance existing features.

Refactoring is mostly applied in perfective software evolution, though it also affects corrective and adaptive evolution. First, well-organized and flexible software allows one to quickly isolate and correct errors. Second, such software ensures that new functionality can be easily added to address changing user requirements.

A known issue about refactoring is automatization. Small steps of refactoring have usually been performed manually, using primitive tools such as text editors with search and replace functionality. This situation eventually leads to corrupt the design of source code, mostly due to the fact that manual refactoring is tedious and prone to errors [2]. Although the choice of which refactoring to apply is naturally made by human, automatic execution of refactorings might result in a major improvement in productivity.

In addition, concerning behavior preservation, TDD informally guides refactoring assisted by unit tests, increasing the correctness of a sequence of transformations. Furthermore, verification of object-oriented programs is highly nontrivial. A number of recent research initiatives have pointed out

directions for formally justifying refactorings. In Opdyke's work, preconditions for refactorings are analyzed [10], whereas Robert's work formalizes the effect of refactorings in terms of postconditions, in order to build efficient refactoring tools [2]. In contrast, Mens [9] apply graph representation to aspects that should be preserved and graph rewriting rules as a formal specification for refactorings.

IV. CAUSES OF REFACTORING

In computer programming, code smell is any symptom in the source code of a program that possibly indicates a problem at steep level.

Often the deeper problem hinted by a code smell can be uncovered when the code is subjected to a short feedback cycle where it is refactored in small, controlled steps, and the resulting design is examined to assist the needs of more refactoring. From the programer's point of view, code smells are forecast to refactor, and what specific refactoring techniques are to be used. Thus, a code smell is a driver for refactoring. Code smell hint that provides can be improved in some where in your code.

Determining a code smell is often a subjective judgment and will often vary by language, developer and its methodology. There are certain tools, such as Checkstyle, PMD and FindBugs for Java, to automatically evaluate for certain kinds of code smells.

When to apply refactorings to the test code, is different from refactoring production code and the test code has a distinct set of smells dealing with the test cases are organized, to study its implementation and interaction with each other. Moreover, improving test code involves a mixture of refactorings from specialized to test code improvements as well as a set of additional refactorings involving the modification of test classes, ways of grouping test cases, and so on [5].

Refactoring (to Patterns)

- Simple Design -> Code Smell -> Refactor
- Refactoring (to Patterns) is the ability to transform a "Code Smell" into a positive design pattern

Following are the examples of some of the Bad Code Smells that are encountered in case (unit/class) design

- Duplicated Code
- · Methods too big
- Nested "if" statements
- Classes with too many instance variables
- · Classes with too much code
- Strikingly similar subclasses
- Too many private (or protected) methods
- Similar looking code sections

- · Dependency cycles
- Passing Nulls To Constructors
- Classes with too little code

V. TEST CASE CODE SMELLS

This section gives an overview of bad code smells that are specific for test code.

A. Self Contained

When a test uses external resources, such as file containing test data, the test is no longer self contained. Consequently, there is no enough information to understand the test functionality, to use it as test documentation.

Moreover, external resources introduces hidden dependencies: if some force mutates such a resource, tests start failing. Chances for this increase becomes more when more tests use the same resource. The use of external resources can be thus eliminated using refactoring Intregral Resource.

B. Resource Optimism

Test code that makes optimistic assumptions about the existence (or absence) and state of external resources (such as particular directories or database tables) can cause nondeterministic behavior in test outcomes. The situation where tests run fine at one time and fail miserably at the other time needs to be avoided. Resource Allocation refactoring used to allocate and/or initialize all resources that are to be used.

C. Resource Interface

Such wars arise when the tests execute you are the only one testing which fails when more programmers run them. This is most likely caused by Resource Interference: some tests in your suite allocate resources such as temporary files that are also used by others. Identified Uniquely is one of the test code refactoring method used to overcome Resource Interference.

D. Setup Method

In the JUnit framework a programmer can write a setUp method that can be executed before each test method to create a fixture for the tests to run. Things start to smell when the setUp fixture is too general and different tests only access part of the fixture. Such setUps are harder to read and understand.

Moreover, they may make tests run more slowly (because they do unnecessary work). The danger of having tests that take too much time to complete is that testing starts interfering with the rest of the programming process and programmers eventually may not run the tests at all.

E. Splitting Method

When a test method checks methods of the object to be tested, it is hard to read and understand, and therefore more difficult to use as documentation. Moreover, it makes tests more dependent on each other and harder to maintain. The solution is simple: separate the test code into test methods that test only one method. Note that splitting into smaller methods

which can slow down the tests due to increased setup/teardown overhead.

F. Assertion Roulette

"Guess what's wrong?" This smell comes from having a number of assertions in a test method that have no explanation. If one of the assertions fails, it becomes difficult to know the cause of concern. Use Asertion Explanation to remove this smell.

G. Class-to-be-tested

A test class is supposed to test its counterpart in the production code. It starts to smell when a test class contains methods that actually perform tests on other objects (for example because there are references to them in the class-to-be-tested). The smell which arises also indicates the problems with data hiding in the production code. Note that opinions differ on indirect testing. Some people do not consider it a smell but a way to guard tests against changes in the "lower" classes. We feel that there are more losses than gains to this approach: It is much harder to test anything that can break in an object from a higher level. Moreover, understanding and debugging indirect tests is much harder.

H. Duplication across Test Class

Test code may contain undesirable duplication. In particular the parts that set up test fixtures are susceptible to this problem. Solutions are similar to those for normal code duplication as described by Fowler [5]. The most common case for test code will be duplication of code in the same test class. For duplication across test classes, it may prove helpful to mirror the class hierarchy of the production code into the test class hierarchy. A word of caution however can introduce dependencies between tests moving duplicated code from two separate classes to a common class.

A special case of code duplication is test implication: test A and B cover the same production code and A fails if and only if B fails. A typical example occurs when the production code gets refactored before such refactoring.

VI. TEST CODE REFACTORING

Bad smell seems to arise more often in production code than in test code. The main reason for this is that, production code is adopted and refactored more frequently allowing these smells to escape.

One should not, however, underestimate the importance of having fresh test code. Especially when new programmers are added to the team or when complex refactorings need to be performed clear test code is invaluable. To maintain this freshness, test code also needs to be refactored. We define test refactorings as changes (transformations) of test code that: (1) do not add or remove test cases, and (2) make test code better understandable/readable and/or maintainable. The production code can be used as a (simple) test case for the refactoring: If a test for a piece of code succeeds before the test refactoring, it should also succeed after the refactoring. This, obviously also means that you should not modify production code while refactoring test code (similar to not changing tests when refactoring production code).

While working on our test code, the following refactorings are encountered:

A. Integral Resource

To remove the dependency between a test method and some external resource, we incorporate the resource in the test code. This is done by setting up a fixture in the test code that holds the same contents as the resource. This fixture is then can be used instead of the resource to run the test. A simple example of this refactoring is to put the contents of a file that is used into some string in test code.

B. Resource Allocation

If it is necessary for a test to rely on external resources, such as directories, databases or files, make sure the test that uses them explicitly creates or allocates these resources before testing and releases them when done (take precautions to ensure the resource is also released when tests fail).

C. Identified Uniquely

Lot of problems originate from the use of overlapping resource names; either between different tests run done by the same user or between simultaneous test runs done by different users. Such problems can easily be overcome using unique identifiers for all resources that are allocated, such as including a time-stamp. When you also include the name of the test responsible for allocating the resource in this identifier, you will have less problems finding tests that do not properly release their resources.

D. Minimize Data

Minimize the data that is setup in fixtures to bare essentials. This will have two advantages: (1) in making them better suitable for documentation and consequently (2) the tests will be less sensitive to changes.

E. Assertion Explanation

Assertions in the JUnit framework have an optional first argument to give an explanatory message to the user when the assertion fails. Testing becomes much easier when you use this message to distinguish between different assertions that occur in the same test. May be this argument should not have been optional.

F. Add Equality Method

If an object structure needs to be checked for equality in tests, an implementation for the "equals" method for the object's class needs to be added. You then can rewrite the tests that use string equality to use this method. If an expected test value is only represented as a string, explicitly construct an object containing the expected value and use the new equals method to compare it to the actually computed object.

VII. CONCLUSION

The large refactoring can improve overall quality of a test case using these set of smells choices. The only concern needs to be understand the selection of refactoring choices. But which refactoring choices should be implemented? We advocates program slicing in conjunction with code smell to guide refactoring process. By slicing the software system one or more bad smells, different refactoring options can examined and evaluated using these sets of smells. Thus the combination of program slicing and set of code smells guides the refactoring process.

A software system essentially needs the refactoring systems for its better performance. Thus this refactoring process assist in its high quality and can prove to be more maintainable techniques. This refactoring process thus can be executed in lower error rates, fewer test cases per module and to increase over all understandability and maintainability in return. In both the design and maintenance phases, these advantages can be realized almost immediately.

REFERENCES

[1] Cohen M. B., P. B. Gibbons, W. B. Mugridge and C. J. Colbourn, Constructing Test Suites for Interaction Testing. 25th International Conference on Software Engineering (ICSE'30), Portland, Oregon, United States, IEEE Computer Society, 2003,38-49.

- [2] Don Roberts. Practical Analysis for Refactoring. (PhD thesis, University of Illinois at Urbana-Champaign, 1999).
- [3] Elbaum S., A. G. Malishevsky and G. Rothermel, Prioritizing Test Cases for Regression Testing, ACM SIGSOFT International Symposium on Software Testing and Analysis, Portland, Oregon, United States, ACM Press, 2000,102-112.
- [4] Fodeh John A. and Niels B. Svendsen, Release Metrics: When to Stop Testing with a clear conscience, Journal of Software Testing Professionals, March 2002.
- [5] Fowler M. Refactoring: Improving the Design of Ex-isting Code (Addison-Wesley, 1999).
- [6] Griswold William G. Program Restructuring as an Aid to Software Maintenance, (PhD thesis, University of Washington, 1991).
- [7] Judson Sheena R., Doris L. Carver, and Robert France. A Metamodeling Approach to Model Refactoring. Submitted to UML' 2003.
- [8] Kang B.-K. and J. M. Bieman. A Quantitive Framework for Software Restructuring. Journal of Software Maintenance, 11, 1999, 245-284.
- [9] Mens Tom, Serge Demeyer, and Dirk Janssens. Formalising Behaviour Preserving Program Transformations. In Proceedings of the First International Conference on Graph Transformation, Springer-Verlag, 2002,286-301.
- [10] Opdyke William, Refactoring Object-Oriented Frameworks, (PhD thesis, University of Illinois at Urbana-Champaign, 1992).
- [11] Robert S. Arnold. Software Restructuring. Proceedings of the IEEE, 77(4), April 1989,607-617.
- [12] Volokh Eugene, VESOFT (1990), Automated Testing.... When and How, (Interact Magazine) .

Extraction of Information from Images using Dewrapping Techniques

Khalid Nazim S. A., Research Scholar, Singhania University, Rajasthan, India. Dr. M.B. Sanjay Pande, Professor and Head, Department of Computer Science & Engineering, VVIET, Mysore, India.

Abstract-An image containing textual information is called a document image. The textual information in document images is useful in areas like vehicle number plate reading, passport reading and cargo container reading and so on. Thus extracting useful textual information in the document image plays an important role in many applications. One of the major challenges in camera document analysis is to deal with the wrap and perspective distortions. In spite of the prevalence of dewrapping techniques, there is no standard efficient algorithm for the performance evaluation that concentrates on visualization.

Wrapping is a common appearance document image before recognition. In order to capture the document images a mobile camera of 2megapixel resolution is used. A database is developed with variations in background, size and colour along with wrapped images, blurred and clean images. This database will be explored and text extraction from those document images is performed. In case of wrapped images no efficient dewrapping techniques have been implemented till date. Thus extracting the text from the wrapped images is done by maintaining a suitable template database. Further, the extracted text from the wrapped or other document images will be converted into an editable form such as

Notepad or MS word document. The experimental results were corroborated on various objects of database.

Keywords: Dewrapping, Template Database, Text Extraction.

I. Introduction

An image may be defined as a two dimensional function f(x, y), where x and y are spatial co-ordinates and the amplitude of f at any pair of co-ordinates(x, y) is the intensity or gray level of the image at that point. When x, y and the intensity values of f are all finite, the digital image is composed of finite number of elements where each has a particular location and value. These elements are called picture elements, image elements, pels and pixels [7][14].Image processing can be broadly categorized into two classes. The first category takes images as input and gives the images as output. The other category takes images as input and gives the attributes of images as output. The entire processing can be listed as: (i). Image enhancement-It involves manipulating an image so that the result is more suitable than original for processing.

- (ii). *Image restoration* It involves improving the appearance of an image based on mathematical or probabilistic model of image degradation.
- (iii). *Colour image processing* Colour can be used as factor or basis for extracting features

of interest in an image.

- (iv). Compression- This reduces the storage required to save an image or bandwidth to transmit an image.
- (v). Morphological image processing- It deals with the tools for extracting image components that are useful in the representation and description of shape.
- (vi). Segmentation- It deals with the partitioning of an image into constituent parts namely autonomous and rugged segmentation [7] [4].

A document is a bounded physical or digital representation of a body of information with capacity (and usually intent) to communicate. Document image processing and understanding has been extensively studied over the past 40 years that has carved a niche out of the more general problem of computer vision because of its pseudo binary nature and the regularity of the patterns used as a "visual" representation of language. In the early 1960s, optical character recognition was taken as one of the first clear applications of pattern recognition and today, for some simple tasks with clean and wellformed data document analysis is viewed as a solved problem. Unfortunately, these simple tasks do not represent the most common needs of the users of document image analysis. The challenges of complex content and layout, noisy data and variations in font and style presentation keep the field active.

Traditionally, document images are scanned from pseudo binary hardcopy paper manuscripts with a flatbed, sheet-fed, or mounted imaging device. Recently, the community has seen an increased interest in adapting digital cameras to tasks related to document image analysis. Digital camcorders, digital cameras, PCcams, PDA's (personal digital assistant) and even cell phone cameras are becoming increasingly popular and they have shown potential as alternative imaging devices.

Although they cannot replace scanners, they are small, light, easily integrated with various networks and more suitable for many document capturing tasks in less constrained environments. These advantages are leading to a natural extension of the document processing community where cameras are used to image hardcopy documents or natural scenes containing textual content [12].

Cameras in an uncontrolled environment have triggered a lot of interest in the research community over the last few years and many approaches have been proposed. However, there has been no satisfactory work presented for dewrapping techniques so far. Wrapping is a common appearance in camera captured document images [13]. It is the primary factor that makes such kind of document images hard to be recognized. Therefore it is necessary to restore wrapped document image before recognition. The documents captured from cameras often suffer from various distortions, like non-planar (wrapped) shape, uneven light shading, motion blur, perspective distortion, under-exposure and over-exposure. But current Optical Character Recognition (OCR) systems do not deal with these distortions when applied directly to wrapped camera-captured document images.

Images when captured will suffer from distortions such as noise, blur and so on. In order to perform operations on document the distortions have to be removed. Noise removal and blur removal is done using filters. There are several types of filters available among them the Gaussian filter is the most efficient filter. Gaussian filters are a class of linear smoothing filters with the weights chosen according to the shape of the Gaussian function. The Gaussian smoothing filter is a very good filter for removing the noise drawn from a normal distribution. Gaussian functions rotationally symmetric in two dimensions i.e. the amount of smoothing performed by the filter is the same in all directions. In image sharpening the goal is to highlight fine details in an image. That is, to enhance details that have been blurred. Fine details in the frequency domain correspond to high

frequencies, thus the use of high-pass filters for image sharpening [3] [10].

Text detection refers to the determination of the presence of text in a given frame (normally text detection is used for a sequence of images). Text localization is the process of determining the location of text in the image and generating bounding boxes around the text [2][7]. Text tracking is performed to reduce the processing time for text localization and to maintain the integrity of position across adjacent frames. Although the precise location of text in an image can be indicated by bounding boxes, the text still needs to be segmented from the background to facilitate its recognition. This means that the extracted text image has to be converted to a binary image and enhanced image is then used for text extraction. Text extraction is the stage where the text components are segmented from the background, enhancement of the extracted text components is required because the text region usually has a low-resolution and is prone to noise.

II. LITERATURE SURVEY

Jian Liang, et.al. proposed a method that is focused on analyzing text and documents captured by a camera which is known as camera-based analysis of text and documents. Camera based document analysis is more flexible to provide capability to capture information for visual communication, indexing, reading graphical text in web pages. In camera based analysis of text and documents, sources of images used are paper based, printed handwritten documents, journal etc. Scanner based process provides good reference and starting point, but they cannot be used directly on camera-captured images.

Content in an image can be perceptual

or semantic content but the text within an image is of more interest as it describes the contents of the image. It can be easily extracted compared to the semantic contents. A variety of approaches to Text Information Extraction(TIE) from images and videos have been proposed for specific applications including page segmentation, address block location, number plate location and content based image or video indexing. Text extraction system has various applications such as portable computers, content based video/document coding, license plate recognition and video content analysis. To enhance performance of text information system it is advantageous to merge various sources as proposed by Keechul et.al.

Portable digital cameras are now used for digitalizing documents and as a fast way to acquire document images taking advantage of their low weight, portability, low cost, small dimensions etc. Several specific problems arise in this digitization process. Rafael et.al, addressed the inherent problems of document image digitization using portable camera. Their work was based on an issue that documents make use of translucent paper in such a way that back-to-front interference was not observed. Also when a document image is taken from the camera the strobe flash causes an uneven illumination of the document. Marginal noise, not only drops the quality of resulting image for CRT screen visualization, but also consumes space for storage and large amounts of toner for printing, which alters the segmentation algorithm of the optical character recognition and thus affects the response obtained in the number of characters and words correctly transcribed. It assumes that the background may be of any colour or texture, provided that there is a colour difference of at least 32 levels between the image background and at least one of the RGB components of the most frequent colour of the document background (paper). Two different experiments were set to evaluate lens distortion. The first one is visual inspection by humans, while the second one is based on analyzing the effect of the compensation of lens distortion in optical character recognition.

monochromatic In images iconographic or artistic value saves storage space and bandwidth in network transmission. The binarized preprocessed documents lowered the number of character substitution and it has also raised the incidence of insertion errors due to spurious noise inserted in the image. This paper presents ways significantly improve the visualization of images whenever displayed on screen of CRT or LCD or printed.

Celine and Bernard presented the solution which is independent of scenes, colours, lighting and all various conditions. Their algorithm was based on multi-hypothesis text extraction.

Yu Zhanga et.al. **proposed** an algorithm based on binary document images which considers the horizontal text that is mostly present in both Arabic and Chinese characters. Wrapped document images should have text lines with a main direction of horizon. Thus several pairs of key points when mapped using Thin Plate splines(TPS) will restore the original image based on an interpolation algorithm[5][8].

Syed Saqib Bukhari et.al. used a novel dewrapping approach based on curled text-lines information, which was extracted using ridges based modified active contour model (coupled snakes). This dewrapping technique is less sensitive with different

direction of curl and variable line spacing. The optical character recognition error rate from wrapped to dewrapped documents was reduced from 5.15% to 1.92% for the dataset collected. approaches for document dewrapping can be divided into two main categories based on the document capturing methodology: one in which specialized hardware arrangement like stereo camera is required for 3D shape reconstruction of wrapped document and the other approach in which dewrapping method is designed for image that is captured using single hand-held camera in an uncontrolled environment.

Faisal Shafait presents an overview of the approaches based on evaluation measure and the dataset used. The methods used are continuous skeletal image representation for document image dewrapping, Segmentation based document image dewrapping, Coordinate Transform Model (CTM) and document rectification for book dewrapping. Dewrapping of documents captured from hand-held cameras has triggered a lot of interest and thus many approaches have been proposed to achieve that. However, there has been no comparative evaluation of different dewrapping techniques so far. A dataset of 102 documents captured with a hand-held camera were created and made freely available online. A text-line, text-zone, and ASCII text groundtruth for the documents in this dataset were made. The results showed that the CTM presented by Wenxin Li et al. performed better than the other two methods, but the difference was not statistically significant. Overall, all participating methods worked well and the mean edit distance was less than 1% for each of them.

Based on the literature survey our main aim is to recover document images.

Hence the proposed technique was applied on gray scale document images and is based on several distinct steps like an adaptive document image binarization, a text line and word detection, a draft binary image dewrapping based on word rotation and shifting and finally a complete restoration of the original gray scale wrapped image guided by the binary dewrapping. The problems encountered are background removal, skew often found in the image in relation to the photograph axes, as documents have no fixed mechanical support in the document image.

A. DATA SAMPLE DESCRIPTION

Text data present in images and video contain useful information for automatic annotation, indexing and structuring of images. Extraction of this information involves detection, localization, tracking, extraction, enhancement and recognition of the text from a given image. The data samples for text extraction were captured as image from a 2Megapixel mobile camera with a resolution of (960 X 1280). The size of each image captured varies between (60 – 80KB). The variations of text present in the images captured will be due to variations in size, style, orientation and alignment followed with a low image contrast and complex background.

The images that were captured are manually classified/ categorized into blur, clean and wrapped images as shown in Fig 1.

Blur Images	Clean Images	Wrapped images
22	ARE	TEXT
OR	HERE	12246
50	HEAVY METAL	12343

Figure 1: Blur Images, Clean Images and Wrapped images

III. IMPLEMENTATION

The process is divided into three main phases namely preprocessing phase, dewrapping phase and the text extraction phase. In preprocessing phase, the quality of an image is enhanced. In dewrapping phase, the wrapped document images (the images which are captured from the cylindrical objects surface) are dewrapped. In the

text extraction phase, the text in the document image is detected, localized and finally extracted into editable form.

1. Pre-processing Phase: The experimental setup for capturing an image containing text requires a 2Megapixel mobile camera which is placed at a standard distance. The measuring scale is used to measure the distance of an

image from the mobile camera. Further, the intensity of light is also varied. The images are captured in a controlled environment at the distances of 10cm, 13cm and 15cm respectively.

- **1.1 Preprocessing with Different Image Documents:** Preprocessing involves several steps which are presented in the following section. In preprocessing we tend to remove noise, blur operation and sharpening as shown in Figure 2.
- 2. Dewrapping Phase: Dewrapping is a two-step approach at the first step, a coarse dewrapping is accomplished with the help of a transformation model that maps the projection of a curved surface to a 2D rectangular area. The projection of the curved surface is delimited by the two curved lines which fits the top and bottom text lines along with the two straight lines that fit to the left and right

text boundaries. Towards the second step, a fine dewrapping is achieved based on words detected. All words pose as normalized guide lines by the lower and upper word baselines.

3. Text extraction Phase: Text data present in images and video contains useful information for automatic annotation, indexing and structuring of images. Extraction of this information involves detection, localization, tracking, extraction, enhancement and recognition of the text from a given image.

However variations of text due to differences in size, style, orientation and alignment as well as low image contrast and complex background make the problem of automatic text extraction extremely challenging.

Pre-processing Phases			
Converting the input RGB image into grayscale image: I = rgb2gray(RGB)	Converting gray image to binary image: BW = im2bw(I, level)	Blur removal: h = fspecial(type) B = filter(A,H)	Sharpening: h = fspecial('unsharp', alpha)
MP65	MP65	311	MP65
MP65	MP65	311	MP65

Figure 2: Block diagram of preprocessing

V. ALGORITHM

Stage 1: Convert the input RGB image into grayscale image:

(The RGB values are normalized to a single gray scale value. Grayscale images are distinct from one-bit black-and-white images, which in the context of computer imaging are images with only two colours black, and white [bi-level images]).

Stage 2: Convert grayscale image into binary image:

Stage 2.1: BW = im2bw (I, level)

The output image BW replaces all pixels in the input image with luminance greater than level with the value 1 (white) and replaces all other pixels with the value 0 (black).

Stage 3: Blur removal:

Stage 3.1: h = fspecial (type)

(Creates a two-dimensional filter h of the specified type. fspecial returns has a correlation kernel, which is the appropriate form to use with imfilter.type is a string having one of the following values: average, disk, gaussian)

Stage 3.2: $\mathbf{B} = \mathbf{imfilter}(\mathbf{A}, \mathbf{H})$

(Filters the multidimensional array A with the multidimensional filter H.)

Stage 4: Sharpening:

Stage 4.1: **h = fspecial ('unsharp', alpha)**

(Returns a 3/3 unsharp contrast enhancement filter. fspecial creates the unsharp filter from the negative of the Laplacian filter with parameter alpha. where alpha controls the shape of the Laplacian and must be in the range (0.0 to 1.0). The default value for alpha is 0.2.)

Stage 5: Dewrapping:

Based on the background of the image, different approaches have been proposed for document image dewrapping. These approaches can be divided into two main categories based on the document capturing methodology: (i) approaches in which specialized hardware arrangement like stereo camera, is required for 3D shape reconstruction of wrapped document and (ii) approaches in which dewrapping method is designed for image that is captured by using single hand-held camera/Mobile in an uncontrolled environment. In our present work, we deal with the second approach using Mobile phones where documents are captured using mobile on curved surfaces [8].

VII. SPECIAL CASES

We have also worked with certain typical special cases where recognition of character was not highly relevant to the one present in the given document image which is mainly due to the style in which characters are represented in the given document image as illustrated in figure 3.

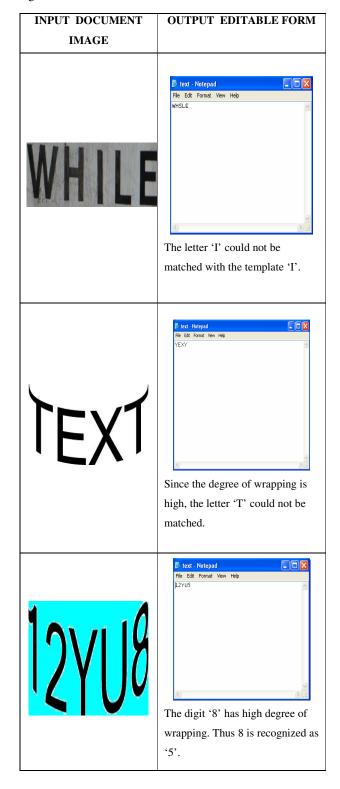
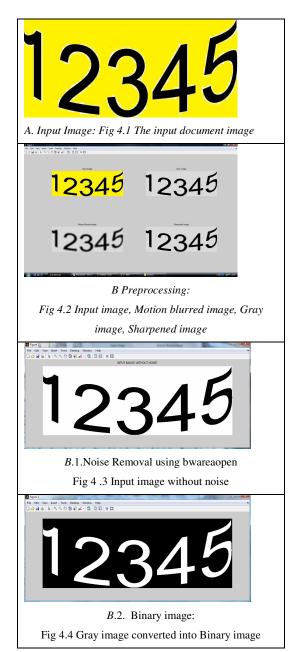


Figure 3: Typical Special Cases

VIII. EXPERIMENTAL RESULTS

Typical results on the various operations of the proposed approach are as shown in Fig.4.



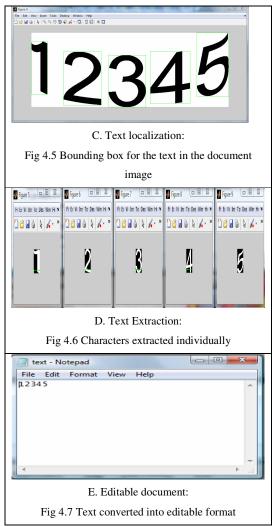


Figure 4: Various Operations performed

IX. CONCLUSION

Image processing is a method that has wide applications in disciplines related to a researcher's preview. In the present work, we have taken up an interesting concept of document image analysis in a broader sense but we have restricted ourselves to text wrapping.

In specific, document analysis plays a very important role in image processing since any information present that has to be

authenticated will be in a form of document only. Thus in the present problem, we have used a concept of template matching to evaluate the text present in an input image with database of characters that have been taken as a knowledge during the process of training.

Image wrapping or dewrapping may be implemented using texture mapping by defining a correspondence between a uniform polygonal mesh and a wrapped mesh. The points of the wrapped mesh are assigned the corresponding texture coordinates of the uniform mesh and the mesh is texture mapped with the original image. Using this technique, simple transformations such as zoom, rotation or shearing can be efficiently implemented.

This paper provides various tools to play with images and also opens up new avenues in the areas such as text extraction and detection. Typically the approach is focused to convert the input document image, that is text (which might be a text present in an image) into its gray image constituents. Before performing any operations the present paper takes into consideration the various preprocessing stages such as blur, cleaning and sharpening to cognize a knowledge base. The algorithm designed for the present work is able to identify the alphanumeric text into its relevant values i.e. its alphabets (capital A-Z & small a-z) and numerals (0-9) that can have the properties of skew also.

REFERENCES

- [1] Syed Saqib Bukhari, Faisal Shafait, Thomas M. Breuel, "Ridges Based Curled Text line Region detection from Gray scale Camera-Captured document Images," 13th Int. Conf. on Computer analysis of Images and Patterns, CAIP'09, Munster, Germany, 2009.
- [2] Fabio Caccia, Roberto Marmo, Luca Lombardi., "License Plate Detection and Character Recognition," In Pasquale Foggia, Carlo Sansone, Mario Vento, editors, Image Analysis and Processing - ICIAP 2009, 15th International

- Conference, Vietri sul Mare, Italy, 2009, proceedings. Volume 5716 of Lecture Notes in Computer Science, pages 471-480, Springer, 2009.
- [3] B. Gatos, I. Pratikakis, K. Kepene, S.J. Perantonis, "Text detection in indoor/outdoor scene images," in: Proc. 16th IEEE international conference on image processing, pp: 127-132, 2009.
- [4] S Jayaraman, S Esakkirajan, T VeeraKumar, "Digital Image Processing," ISBN: 0070144796, McGraw-Hill Education, India, 2009.
- [5] Celine Mancas-Thillou, Bernard Gosselin "Natural Scene Text Understanding. Computer Vision and Image Understanding," Volume 107, Issue 1-2, pp: 97-107, 2007.
- [6] Stephen J.Chapman,"Mat lab Programming for Engineers", 3rd edition, McGraw-Hill Education, 2007.
- [7] Rafael C. Gonzalez, Richard E. Woods, "Digital Image processing, Publishing House of Electronics-An Approach-Effect of an Exponential Distribution on different Processing", Second edition, 2005.
- [8] Aleksander Recnik, Gunter Mobus, Saso Sturm,"IMAGE-WRAPP: A real-space restoration method for high-resolution STEM images using quantitative HRTEM analysis," Elsevier, 2005.
- [9] Wenxin Li, Jane You, David Zhang: "Texture-based palm print retrieval using a layered search scheme for personal identification," IEEE Transactions on Multimedia, Volume 7, Number 5, pp: 891-898, 2005
- [10] Yu Zhang, Shie Qian and Thayananthan Thayaparan," Two new approaches for detecting a maneuvering air target in strong sea-clutter," Radar Conference, 2005 IEEE International, pp. 83-88, 2005.
- [11] Keechul Jung, Kwang in Kim, Anil K. Jain,"Text Information Extraction in Images and Video: A Survey," Pattern recognition, volume37, issue 5, pp: 977-997, 2004.
- [12] Jian Liang, David Doermann and Huiping Li." Camera-Based Analysis of Text and Documents: A Survey," International Journal on Document Analysis and Recognition, Volume 7(2+3), pp: 83 --104, Springer-Verlag 2005.
- [13] Faisal Shafait, Thomas M. Breuel, "Document Image Dewrapping Contest," In proceedings of 17th ICPR, Volume 1, pp 482-485, 2004.
- [14] Rafael C. Gonzalez, Richard E. Woods, StevenL.Eddins, "Digital Image Processing UsingMATLAB," Prentice Hall, Pearson education, 2004.

SECURED AUTHENTICATION PROTOCOL

SYSTEM USING IMAGES

G. Arumugam

Prof. & Head, Department of Computer Science Madurai Kamaraj University Madurai, India.

Abstract—In order to protect secret information from sensitive and various applications, secured authentication system should be incorporated; it should contain security and confidentiality. Even if it is assumed that the cryptographic primitives are perfect, the security goals may not be achieved: the system itself may have weaknesses that can be exploited by an attacker in network attacks. In this paper a Secured Authentication Protocol System using Images (SAPSI) is presented. It ensures confidentiality, and authentication using server and Image based authentication mechanism.

Keywords- Confidentiality, Security, Server, Image-Based Authentication System, Authentication.

I. INTRODUCTION

A significant challenge in providing an effective network system defence mechanism is to detect the intrusions and implement counter-measures. Organizations who use Secured Authentication system tolerate no leakage at all. Cryptographic primitives are useful tools but security of the primitives does not guarantee security of the system. Usage of different level of security provides a security policy that allows the classification of data and users based on a system of hierarchical security levels combined with a system of non-hierarchical security categories.[1, 5, 6].

Cryptographic mechanisms are communication systems that rely upon cryptography to provide security services across distributed systems. Applications increasingly rely on encryption services provided by cryptographic systems to ensure confidentiality and authentication during secure transactions over the network. However the security provided by these encryption services might be undermined if the underlying security system has any flaws in the design or implementation. Weaknesses in security systems such as misuse of encryption, compromising the private encryption key etc., are yet to be addressed. [8].

Secured Authentication System is an application of a computer system to process information with different sensitivities (i.e. classification of information at different levels) to permit simultaneous access by users with different security clearance and to prevent users from obtaining access to information for which they lack authorization. Secured Authentication has two goals: first goal is to prevent unauthorized personnel from accessing information. Second

R. Sujatha

Research Associate, SSE Project, Department of Computer Science Madurai Kamaraj University Madurai, India.

goal is to prevent unauthorized personnel from declassifying information. The traditional view of secured authentication is one of ensuring that information at a high security classification cannot flow down to a lower security classification.[1, 3, 12].

In this paper, Secured Authentication Protocol System using Images is proposed. It overcomes the identified drawbacks of existing systems. The attacks on existing model embedded in encrypted sessions are detected as monitoring the processes taking part in the systems is integrated. The new system uses encryption mechanisms. Hence the inside information is protected and also the outside attacks are prevented. To establish this, a server with authentication mechanism is used. Types of attacks were proscribed in the proposed system are Brute force attack, Dictionary attack, Keyloggers, Shoulder Surfing, Man-In-The-Middle attack and Database Server Compromise attack.

Brute force attack. The hacker can try two kinds of Brute force attacks on this system. One is re-using of images and another is without re-use of images. For a user, there will be a unique password of length 8 or above selected in SAPSI for the given session. Possible image patterns were dynamically changed on every session along with random numbers. By performing this attack in SAPSI hacker unable to break the password because it needs two processes.

Dictionary attack. Dictionary attack is one of the most commonly used techniques to break a Password-based system. If same kind of sequences appeared in the network for a long time it can be guessed by the hacker.

Keyloggers. Keylogger is a program, which captures the user's keystrokes and sends this information to the hacker. The natural protection for an authentication system from the keylogger is to have a one-time password (or Dynamic password).

Shoulder Surfing. Shoulder surfing is looking over someone's shoulder when they enter a password or a PIN code. It is an effective way to get information in crowded places because it is relatively easy to stand next to someone and watch as they fill out a form, enter a PIN number at an ATM machine, or use a calling card at a public pay phone. Shoulder surfing can also be done at a distance with the aid of

binoculars or other vision-enhancing devices to know the password.

Man-In-The-Middle Attack. A man in the middle attack is one in which the attacker intercepts messages in a public key exchange and then retransmits them, substituting his own public key for the requested one, so that the two original parties still appear to be communicating with each other.

This strategy is implemented to protect information from unauthorized disclosure or modification and to provide mechanisms to authenticate users participating in the exchange of information.[7].

In section 2 related works are discussed with their drawbacks.

Section 3 discusses the overview of Proposed Secured Authentication System with server and Authentication mechanism using images methodology.

In section 4 implementation details related to the system are presented. Conclusion is given in section 5.

II. RELATED WORK

Enhanced authentication mechanism using multilevel security model (EAMMSM) is the system that belongs to and applies multilevel security. Any sensitive application it includes confidential and secret information which must be used effectively in complicated and authenticated procedures. Using five levels of authentication methods with a set of privileges assigned, each user has to surpass 50% of every level to get the privileges rights.[1].

During authentication the information was hacked from the network plane using network analyser tool. Leakage of information occurred in three levels while transmitting answers with username and multiple questions methods.

In Improving text password through persuasion (ITPTP), users entered their passwords with visibility.[2].

Users tend to choose their passwords in a simple manner by entering visibility method, which makes the hacker to know with shoulder-surfing process.

An authentication method combining text and graphical passwords (AMCTGP), and users selecting their passwords using random numbers assigned to images, is given in [11].

Users selecting their passwords by clicking random numbers listed in the selection panel can be identified by a hacker using movie-clip camera phones.

In Multiple password interference in text and click-based graphical passwords (MPITCGP), users select their passwords from the given image as pass points.[10].

Users' selecting their passwords from the given image is a hectic process. If any mismatch of pass points occurred the original user itself would be unable to get authentication even by knowing pass point selections.

In Pass Pattern System (PPS): A Pattern-Based User Authentication Scheme, data hacked from database through database compromise server attack is represented. [7].

There are several attempts reported in literature about authentication schemes in lieu of the traditional Password-based system. Each attempt is successful in increasing the strength of the system against some of the known attacks.

They are either computationally intensive or they require additional hardware/software in the infrastructure. In this section we review the current attempts, identify the gaps and emphasize the motivation for developing Secured Authentication Protocol System using Images.

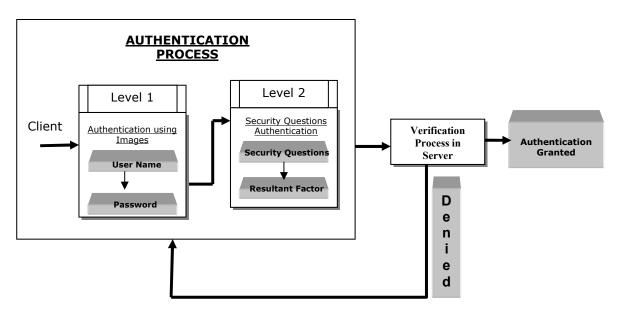


Figure 1: Secured Authentication Protocol System using Images Flow Diagram

Motivation for Secured Authentication Protocol System using Images: We proposed secured authentication system is robust against attacks such as the brute force, shoulder surfing, social engineering, database server compromise attack and Man-In-The-Middle attacks. It incorporates the essence of Image-based authentication system.

III. SECURED AUTHENTICATION PROTOCOL SYSTEM USING IMAGES

This system involves the use of authentication mechanism and a server that minimizes the hacking by the attackers. It monitors the clock cycle process effectively. Two processes are involved in this system. They are a) Authentication using Images and b) Security Questions Authentication using server represented as flow diagram in Figure 1.

A. Authentication using Images

This is a Image-based authentication system based on the premise that 'humans are good at identifying, remembering and recollecting graphical image patterns than text patterns'.[9].

In SAPSI the client gets authenticated in two levels. In the first level the client gets authenticated using username and password method with graphical image patterns. It is illustrated in Figure 2.

For providing the password the client has to enter the index number provided at the images. While entering index numbers in the password area it will be hidden and bullet marks will be displayed. For example, if the client chooses images rose, white lion and lord shiva then the index numbers 27, 44 and 17 should be entered in a selected order. While confirming password images index numbers were shuffled, so user has to re-enter the password by giving different index numbers according to the images chosen. Here both image patterns and index numbers are represented as dynamic arrangements in every login attempt. Due to this setup no one would be able to read or guess the mechanism involved.

For every authentication the images were shuffled and index numbers were varied and shuffled. It is represented in Figure 3.



Figure 2: A sample Secured Authentication Protocol System using Image Patterns

The client has to enter the index numbers according to the selected images in an order given during registration. As per the selection made during registration, the client has to enter index numbers now as 29, 34 and 61.

Each image will be mapped with a corresponding number which is stored in the Image-Map table. Instead of comparing the images, the mapped numbers are compared. It serves as user friendly for the end-user and machine friendly for the system by reducing the comparison time by using numbers rather than images. A mapping mechanism which validates the index numbers with hidden letters is represented in Table I.



Figure 3: A sample shuffling mechanism of Secured Authentication Protocol System using Image Patters.

The client can select the images on some sequences familiar to him/her. Due to shuffling mechanism, this method reduces the guess ability of the persons who are related to the clients. During entry of password, only bullets appear in the password area which avoids the shoulder surfing attacks.

When sending random numbers in the network plane, it will be converted into a computed ascii value, so that Man-In-The-Middle attack is prohibited.

TABLE I
A SAMPLE IMAGE-MAP MECHANISM FOR SAPSI

Image	Const Hid	Random Numbers		
Numbers	Characters	1 Itera- tion	2 Itera- tion	3 Itera- tion
I1	AO	23	15	20
12	IP	70	21	24
I3	LJ	31	10	18
I4	X1	41	16	13
15	YU	12	19	35
16	MK	17	29	26
I7	HR	27	34	90
18	EW	44	61	67
19	SA	55	65	58

Using this mapping mechanism the shuffling process of images and index numbers are generated. The images are validated only by using the hidden characters and index numbers which reduce the time complexity of comparing the images.

The image positions are generated using permutation sequences. Let $A = \{11, 12, 13\}$, this set can be arranged in 3! ways as,

[I1] [I2] [I3]

[I1] [I3] [I2]

[I2] [I1] [I3]

[I2] [I3] [I1]

[13] [11] [12]

[I3] [I2] [I1]

For n images n! Sequences were generated and it will be used randomly for every attempt of registration or login.

Security Potency of Secured Authentication Protocol System using Images:

In general, several attacks are possible on an authentication system. For any authentication system, the hacker can attack at least at three places: they are server, client and the communication link. The attack on server includes Brute force attack, Dictionary attack and compromising the server as a whole. At the client, the possible attacks are key logging and shoulder surfing. Finally on the communication link, the possible attack is Man-In-The-Middle attack, which can be done using packet sniffers.[7].

In terms of the data being passed from the user to the server the data stored in the secured server is comparable with the classical Password-based authentication system. In both cases, user sends the username and a password. This will be compared with the registry in the database. But because of the dynamic nature of password selection system, SAPSI is more secure than ordinary password-based scheme to attacks such as Brute force, Dictionary attack, Keylogger, Shoulder surfing and Server database compromise attack. The best known solution for such attacks is to use cryptography protocols at the server or on the communication link. In this we analyse the impact of the four attacks mentioned here on SAPSI.

On analysing Brute force attack - I in SAPSI, if the hacker wants to guess the password, the probability of success will be $1/(64^4) = 5.96046\text{E}-08$ (Since there are unlimited images, 64 images are taken as sample). If the guess is wrong, probability of success will remain the same for the next guess. It is because the password will change with every attempt. Hence,

The probability of success for every attempt = $1/64^{n}$

The other way of doing Brute force attack - II is to try all combinations of positions. For example, if we consider a 8x8

Image Pattern setup there will be 64^n (if selection of images includes reuse of images) or $^{64}P_n$ (without reuse of images) different images of length n.

Number of possible Image Patterns =
$$\begin{cases} \frac{(N^2)^n}{(N^2-n)!} \\ \frac{(N^2-n)!}{(N^2-n)!} \end{cases}$$

Number of possible Image Patterns for the size of N x N matrix with re-use of images as passwords $(N^2)^n$ is illustrated in Table II and without re-use of images as passwords $(N^2)!/(N^2-n)!$ is represented in Table III.

TABLE II

POSSIBLE RE-USE OF IMAGE PATTERNS

Size of

Length	of the	Image	Password - n
--------	--------	-------	--------------

12	10	8	6	4	Matrix - N
2.81475E+14	1.0995E+12	4294967296	16777216	65536	4
4.73838E+18	3.6562E+15	2.8211E+12	2.177E+09	65536	6
4.72237E+21	1.1529E+18	2.8147E+14	6.872E+10	16777216	8
7.97664E+22	1.2158E+19	1.853E+15	2.824E+11	43046721	9
1E+24	1E+20	1E+16	1E+12	100000000	10
9.84973E+24	6.7275E+20	4.595E+16	3.138E+12	214358881	11
7.94968E+25	3.8338E+21	1.8488E+17	8.916E+12	429981696	12
5.42801E+26	1.9005E+22	6.6542E+17	2.33E+13	815730721	13

TABLE III

POSSIBLE IMAGE PATTERNS WITHOUT RE-USE OF IMAGES

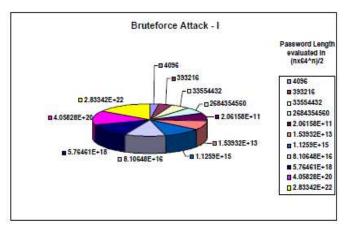
Length of the Image Password - n

Size of Matrix - N	4	6	8	10	12
4	43680	5765760	518918400	2.9059E+10	8.718E+11
6	1413720	1402410240	1.2201E+12	9.2239E+14	5.996E+17
8	15249024	53981544960	1.7846E+14	5.4967E+17	1.573E+21
9	39929760	2.33669E+11	1.2969E+15	6.8163E+18	3.388E+22
10	94109400	8.58278E+11	7.5031E+15	6.2816E+19	5.032E+23
11	203889840	2.76719E+12	3.6278E+16	4.5913E+20	5.606E+24
12	412293024	8.02322E+12	1.5169E+17	2.785E+21	4.963E+25
13	787083024	2.12985E+13	5.6241E+17	1.4488E+22	3.64E+26

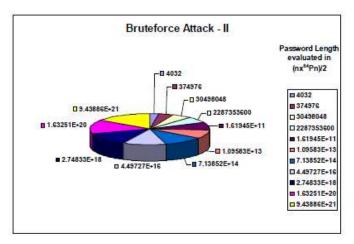
To break the system, the hacker on an average has to break $(nx64^n)/2$ images (with reuse) or $(nx^{64}P_n)/2$ (without reuse).

Number of images that are to be broken =
$$\frac{n (N^2)^n}{2 n N^2!}$$
$$2(N^2 - n)!$$

Number of images that are to be broken by evaluating the length of the password in Brute force attack – I method is depicted in Graph 1 and Brute force attack – II is represented in Graph 2.



Graph 1: Number of Images that are to be broken with reuse of Images.



Graph 2: Number of Images that are to be broken without reuse of Images.

N represents the size of the Image Patterns and n represents the length of the password.

Analysing Dictionary attack in SAPSI, commonly used images with client guess sequences can be possible (if images and random numbers are static). However, here the image pattern changes randomly on every presentation or session; it approaches the behaviour of one-time pad.

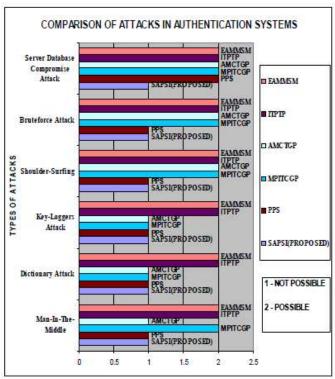
SAPSI, being a dynamic password system, is not vulnerable to keyloggers. Even if the hacker gets the password of the client of a SAPSI system, this password cannot be reused by the hacker to login to the system, because of the dynamic nature of the Image Pattern system.

Shoulder surfing can be done easily on the password system, just by seeing the keys that the user is typing. But to decode the password in SAPSI, the hacker has to see both the key sequence and Image patterns and do a mapping before user submits the page. So shoulder surfing is of little or no use in SAPSI as compared to a password-based system.

In the case of SAPSI, using Man-in-the-middle attack the attacker is not able to get original messages because the images and random numbers changed dynamically on every presentation or session.

Comparing these attacks and it is represented in Graph 3.

The images used for password selection can be of any kind. Depending on the application it can be varied. For sample discussion nature images were used. For implementation characters, numbers and special characters were used as images. Two digit and three digit random numbers were used in implementation. In compact display applications two digit random numbers preferred and in large display applications three digit random numbers preferred to mystify hackers.



Graph 3: Comparison of Attacks in Authentication Systems using Existing and Proposed System.

B. Security Questions Authentication

In second level the client gets authenticated using security questions. A 10-digit number is issued to the client at the time of registration. The client has to answer three security questions and the results are encrypted with a 10-digit number.

A resultant factor is passed over the network plane for validation to the server.

Encryption Process

- Three security questions queried (s1, s2, and s3).
- Ascii value evaluated for two security questions.(a1 and a2)
- Bitwise operation is performed,
 - sum1=(a1 & a2) | s3
- resultant factor (sum2) = sum1 \oplus id
- Ascii value of resultant factor (sum2) send to verifier.

Verification Process

- During Client registration a shared 10-digit key (id) and resultant factor (sum2) issued to server.
- Authentication process: achieved result (sum3) of client ⊕ resultant factor (sum2).
- Authentication granted a shared 10-digit key (id) generated. If not then authentication denied.

The server decrypts the resultant factor and gets the registration number of the client.

After passing the Authentication using Images level and Security questions authentication level, the client gets authenticated.

IV. ANALYSIS AND IMPLEMENTATION

In this new system all drawbacks of existing methods are overcome with new secured authentication protocol system using images. This system is implemented both in single client and multiple clients with server.

Only two levels are used for authentication with single server to authenticate clients. No repetitive methods are used in this proposed method which does not irritate the client. No leakage of information is possible in this new method which avoids the Man-In-The-Middle attack. In [1] leakage of information which occurred during sublevel transitions is avoided in this new system. When entering password no visibility is there which protects the shoulder surfing attacks from related persons. In [2, 11] given passwords are processed using shoulder surfing and if any person tries to hack the password using capture devices, which is protected in new system by giving passwords in a hidden manner. Even if any capturing devices are used to capture images, no one will get information due to hidden bullets in the password area. It is very difficult to remember the pass points in an image [10]. This difficulty is avoided in this new system by selecting the random numbers in the images. This avoids the confused remembrance of pixel positions instead of whole images.

Both existing and proposed systems were implemented and the time difference is evaluated and it is represented in Figure 5. The total image position sequences were generated for 9 images are 362880 and the results were shown in Figure 6.

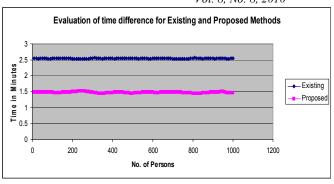
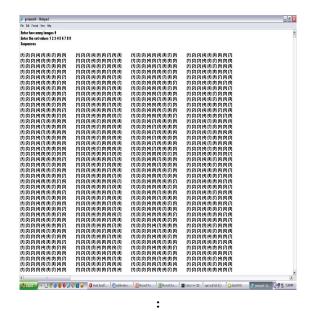


Figure 5: Evaluation of time difference for existing and proposed methods.



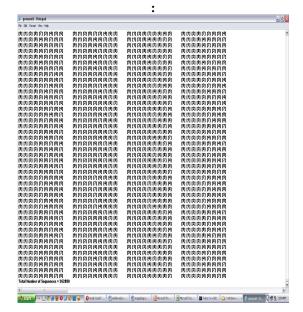


Figure 6: Generated Image Position Sequences for 9 images.

V. CONCLUSION

In order to improve the confidence in security system, detecting intrusions in those systems plays a vital role, as the security system design is not always perfect. Our system overcomes the problem encountered in existing systems and ensures the confidentiality and authentication when sending a message.

ACKNOWLEDGMENT

This paper is part of SSE Project funded through a National Technical Research Organization, New Delhi is gratefully acknowledged.

REFERENCES

- [1] Abdulameer Hussain, "Enhanced Authentication Mechanism Using Multilevel Security Model", Faculty of Science and Information Technology, Zarka Private University, Jordan, International Arab Journal of e-Technology, Vol. 1, No.2, June 2009.
- [2] Alain Forget, Sonia Chiasson, P.C. van Oorschot, Robert Biddle, "Improving text passwords through persuasion", School of Computer Science, Human Oriented Technology Lab, Carleton University, Ottawa, Canada, {aforget, chiasson, pauly}@scs.carleton.ca, robert_biddle@carleton.ca. Symposium on Usable Privacy and Security (SOUPS) 2008, July 23-25, 2008, Pittsburgh, PA, USA.
- [3] Atul Kahate, Cryptography and network security, The Tata Mc-Graw Hill publications.
- [4] Bruice Schneier, Applied Cryptography, Protocols, Alogrithms and Source Code in C, Second Edition, Published by JOHN WILEY and SONS, Reprint 2007.
- [5] Hubert common and vitally shmatikov, Is it possible to decide whether a cryptographic protocol is secure or not?

- [6] Ming-Qing Ling, Wei-Wei Liu, Proceedings of the Seventh International Conference on Machine learning and Cybernetics, Kunming, 12-15 July 2008. Research on IDS based on Levenberg-Marqurdt algorithm.
- [7] T. Rakesh Kumar and S.V. Raghavan, PassPattern System (PPS): A Pattern-Based User Authentication Scheme, NSL, Department of Computer Science and Engineering, IITM, Chennai, India.
- [8] Sachin P. Joglekar, Stephen R. Tate, Protomon: Embedded Monitors for Cryptographic protocol Intrusion Detection and Prevention. Dept. of Computer Science and Engineering, University of North Texas, Denton, TX 76203. {spj0004, srt}@cs.unt.edu.
- [9] R. N. Shepard, C.:Recognition memory for words, sentences and pictures, Journal of verbal Learning and verbal Behavior, vol. 6, pp. 153—163 (1967).
- [10] Sonia Chiasson, Alain Forget, Elizabeth Stobert, P.C. van Oorschot, Robert Biddle, "Multiple Password interference in text and click-based graphical passwords", School of Computer Science, Human Oriented Technology Lab, Carleton University, Ottawa, Canada, {aforget, chiasson, pauly}@scs.carleton.ca, robert_biddle@carleton.ca., estobert@connect.carleton.ca, The definitive version was published in ACM CCS'09 November 9-13, 2009, Chicago, Illinois, USA. Copyright 2009 ACM 978-1-60558-352-5/09/11...\$10.00. http://people.scs.carleton.ca/~paulv/papers/ccs09.pdf.
- [11] P. C. Van OorSchot Tao Wan, "TwoStep: An authentication method combining text and graphical passwords", School of Computer Science, Carleton University, Ottawa, Canada, {paulv, twan}@scs.carleton.ca, E-Techlologies: Innovation in an open world 4th International Conference, MCETECH 2009, Ottawa, Canada, May 4-6, 2009, Proceedings.
- [12] William Stallings, Cryptography and network Security principles and practices, 2006 by pearson education, Inc.

SIP and RSW: A Comparative Evaluation Study

Mahmoud Baklizi¹, Nibras Abdullah¹, Omar Abouabdalla¹, Sima Ahmadpour¹.

1: National Advanced IPv6 Centre of Excellence

1: Universiti Sains Malaysia

1: Penang, Malaysia

Abstract— Voice over internet protocol (VoIP) is a technology that uses Internet to transmit voice digital information. The Session Initiation Protocol (SIP) and Real time Switching (RSW) are signaling protocols that emerged as a new VoIP which gained popularity among VoIP products. In literature, many comparative studies have been conducted to evaluate signaling protocols, but none of them addressed the targeted protocols. In this paper, we make a comparative evaluation and analysis for SIP and RSW using Mean Opinion Score rating (MOS). We found that RSW performs better than SIP under different networks in terms of (packet delays).

Keywords- VoIP; MOS; InterAsterisk eXchange Protocol; Real Time Switching Control Criteria and Session Initiation protocol.

I. Introduction

Nowadays, the use of distributed computer network systems has been increased in many areas of government, academia and industry. Video conferencing system is one of computerbased communication applications. The idea of video conferencing appeared for the first time in the 1920s [1]. The task of video conferencing concentrates on individuals to be together in space and time, and makes groups more effective at their work by applying different services such as telephony service over IP networks that are known as IP telephony or Voice over IP. VoIP communication usually consists of two protocols: (i) Signaling protocols that are used to setup a voice conversation and manage voice sessions (ii) Media transfer protocols that are used for exchanging voice data traffic during one session lifetime. [2]. One of the most important functions in the VoIP infrastructure is signaling session. Signaling session should be by any VoIP protocol before transmitting any media type. Therefore, it allows various network components to communicate with each other to setup and to tear down calls [3]. Recently, there is strong focus on the development of scalable VoIP protocols, such as SIP, RSW and IAX.

Large efforts have been done to study SIP. It was standardized in the IETF (Internet Engineering Task Force) RFC 2543 and further extended in RFC 3261. It is used for creating, modifying, and terminating sessions with one or more participants [5][6].

Session Initiation Protocol has many features: (i) the service of text-based which allows easy implementation in object

oriented programming languages such as Java and Perl. These allow easy debugging, and most importantly make SIP flexible and extensible. (ii) Less signaling. (iii) transport-layer-protocol neutral (iv) parallel search [5][7].

Real time Switching appeared in late 1993 as a control mechanism for multimedia conferencing. It was designed by Network Research Group (NRG) in school of computer science-University Science Malaysia (USM) .The goal of RSW Control Criteria is how to conduct a conference around a meeting table [8][9][10]. Moreover, RSW is used to handle two issues in multimedia conferencing. The first one is to handle the confusion generated while everyone tries to speak at the same time. The second issue is the tremendous amount of network traffic generated by all participating sites [6].

More recently, InterAsterisk exchange protocol has been emerged as new protocol that improved the voice quality. It has also many features such as simplicity, NAT-friendliness, efficiency and robustness [2][3]. There are several goals for this protocol. The main goals of this protocol are (i) Minimizing bandwidth usage for signaling and media transfer, with a particular emphasis on voice (ii) Ensuring NAT transparency (iii) Ability to exchange dial plans (iv) Efficient implementation of intercom and paging features. On the other hand, IAX can be used with different types of streaming media such as video and voice calls [4]. In this paper, we have selected two case studies that depend on IAX protocol. In the first case study, the researchers made comparison between IAX and RSW [3]. While in the second case study the researchers compared between IAX and SIP [2]. According to the last two comparisons, we have made a comparative evaluation of SIP and RSW by using Mean Opinion Score rating (MOS).

II. RELATED WORK

Before transmitting any media, the signaling session, which is the most important function in VoIP, allows different network components to communicate between each other to set up and terminate calls. Mean Opinion Score (MOS) is used as a method to assess call quality. In addition, it is used to measure subjective voice quality. To measure network performance, we use the MOS rating which is the most widely used assessment technique. The listening subjects were given a scale from 1 to

5, where 1 = bad, 2 = poor, 3 = fair, 4 = good, and 5 = excellent[11].

Recently, there are many researchers concentrated on this function and compared between different VoIP protocols. The study conducted by [2] compared the performance between IAX and SIP protocols. They have indicated that both protocols perform comparably in the presence of fixed delay. Therefore IAX appeared to perform slightly better as shown in Figure 1 [2].

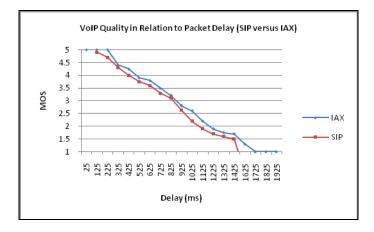


Figure 1: Packet Delay - SIP versus IAX [2]

The researchers in [3] compared the performance between IAX and RSW protocols. They indicated that both protocols perform comparably in the presence of fixed delay. Therefore IAX appeared to perform slightly better than RSW as shown in Figure 2 [3].

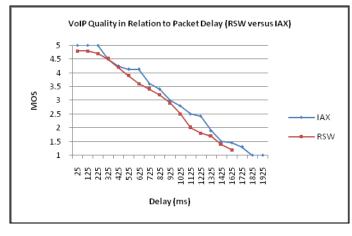


Figure 2: Packet Delay - RSW versus IAX [3]

III. ANALYSIS AND EVALUATION

The predicated evaluation is based on previous two comparisons that were conducted by [2] and [3]. We can see

that the packet delay in IAX is better than that in SIP and RSW. The gained results are extracted using the SPSS statistical tool. The Interclass Correlation Coefficient statistical test is used to analyze IAX with comparison to the previous two studies. As a result, the correlation between the two variables is high. The single measure Interclass correlation coefficient is 0.995, the test value for absolute agreement is 0.99999, and *p* is approximately equal to 1. The IAX protocol regression formula and curve were identical in previous two studies. An indirect comparison between RSW and SIP was made directly in terms of fixed packet delay. Figure 3 indicates that the RSW protocol performs slightly better in the presence of fixed packet delay than SIP protocol.

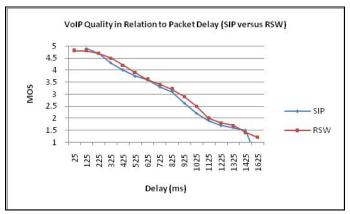


Figure 3: Packet Delay - SIP versus RSW

CONCLUSION

This paper suited the performance of the RSW and SIP in term of packet delay. From the evaluation and analysis, we conclude that RSW achieves more improvement than SIP in VoIP in relation to packet delay via the MOS score. Further research on RSW control criteria and SIP protocols and their performance under various conditions is recommended.

Acknowledgment

We would like to thank Hani Mimi and Abdullah Dahbali for their support and guidance throughout the study and analysis.

REFERENCES

- E. M. Schooler, "Conferencing and collaborative computing," Multimedia Systems. Vol. 4, pp. 210-225, 1996
- [2] T. Abbasi, S. Prasad, N. Seddigh, and I. Lambadaris, "A comparative study of the SIP and IAX VoIP protocols," Electrical and Computer Engineering, Canadian Conference, pp.179-183, 1-4 May 2005
- [3] S. Manjur, M. Mosleh, O. Abouabdalla, W. Tat, and A. Manasrah, "Comparative Evaluation and Analysis of IAX and RSW,"(IJCSIS). Vol. 6,2009

- [4] M. Spencer, F. Miller, "IAX Protocol Description," February 2005, http://www.cornfed.com/iax.pdf.
- [5] I. Dalgic and H. Fang, "Comparison of H.323 and SIP for IP telephony signaling," in Proc. of Photonics East, (Boston, Massachusetts), SPIE, Sept. 1999.
- [6] O. Abouabdalla, R. Sureswaran, "Enable Communications between The RSW Control Criteria and SIP Using R2SP," Distributed Frameworks for Multimedia Applications, 2006. The 2nd International Conference on, vol., no., pp.1-7, May 2006.
- [7] Y. Zhang, "SIP-based VoIP network and its interworking with the PSTN," Electronics & Communication Engineering Journal, December 2002, Volume 16, Issue 6, pg 273-282
- [8] R. Sureswaran, "A Reflector Based System to Support the RSW Multimedia Conferencing Control Criteria," IASTED International Conference on Networks, Orlando, January 1996.
- [9] R. Sureswaran, Subramanian, R.K, H. Guyennet, and M. Trehel, "Using the RSW Control Criteria To Create A Distributed Environment for Multimedia Conferencing," In Proceedings of REDECs '97. Penang, Malaysia. 27-29 November 1997.
- [10] R. Sureswaran, "A Distributed Architecture to support Multimedia Applications Over the Internet and Corporate Intranets," In Proceedings of SEACOMM '98. Penang, Malaysia. 12-14 August 1998.
- [11] http://www.itu.int/rec/T-REC-P.800/en



Mahmoud Khalid Baklizi is a researcher pursuing his PhD in Computer Science at the National Advanced IPv6 Center of Excellence in University Sains Malaysia. He received his first degree in Computer Science from Yarmouk University, Jordan, 2002 and his Master degree in Computer Information System from the Arab Academy for Banking and Financial Sciences, Jordan in 2008. His research area of interest includes Multimedia Networking.



Nibras Abdullah Faqera received his Bachelor of Engineering from College of Engineering and Petroleum, Hadhramout University of science and technology, Yemen, 2003. He obtained his Master of Computer Science from School of Computer Science, Universiti Sains Malaysia in 2010. He is academic staff member in Hodeidah University, Yemen. He is researcher pursuing his PhD in Computer Science at the National Advanced IPv6 Center of Excellence in University Sains Malaysia. His research area

of interest includes Multimedia Conferencing System (MCS).



Dr. Omar Amer Abouabdalla is a senior lecturer and head of the technical department in the National Advanced IPv6 Centre (NAv6) - University Science Malaysia (USM). Dr. Omar is the Chairman of multimedia working group (a sub working groupin APAN), Asia Pacfic Advanced Network (APAN) is a high bandwidth network that will interconnect the Asia Pacific Countries. He is also a member of Internet Engineering Task Force (IETF) and

a member of Editorial Board for Journal of IT in Asia. Dr. Omar is heavily involved in researches carried by NAv6 center, such as Multimedia Conferencing System (MCS) and IPv6 over Fiber project. He has more than five years experience in the field of IPv6 and more than ten years in the field of Multimedia Network.

A role oriented requirements analysis for ERP implementation in health care Organizations

Kirti Pancholi¹, Durgesh Kumar Mishra²

¹Acropolis Institute of Pharmaceutical Education and Research, Indore, MP, India

²Acropolis Institute of Technology and Research, Indore, MP, India

Abstract- Information is worthwhile only if it can be accessed at the right time, by the right person & is useful for the purpose defined. Health care providers have a strong tradition of safeguarding private health information. Today's world belongs to Information Technology. With information broadly held and transmitted electronically, the rule provides clear standards for all parties regarding protection of personal health information. Medical resources integration concerns has also been a long-standing problem, which need to work in collaboration with information technology, aiming at a common goal. The complexity and extension of roles of the planning system demand extensive seamless integrations in the organization. Medical enterprise resources planning (ERP) integrates each level of healthcare staff by providing information and knowledge in timely manner, making the ERP a synchronizing solution for all the roles required in the organization for various timely decision makings. From this motivation, this paper proposes a role-oriented requirement definition analysis for ERP implementations in the organizations. Integrated hospitals need a central planning and control system to plan patients' processes and the required capacity. Given the changes in healthcare one can ask the question what type of information systems can best support these healthcare delivery organizations. We focus in this review on the potential of enterprise resource planning (ERP) systems for healthcare delivery organizations.

Keywords: Patient Logistics, Planning and control, clinical management, ERP, AP.

I. INTRODUCTION

Information technologies in healthcare is more about how it originally developed for manufacturing are fitted to support clinical and administrative work in hospitals. Common denominators for these information technologies are Enterprise resource planning systems or simply ERP. As the healthcare sector faces increasing demands from political and public sides to document increased cost- effectiveness, to make optimal use of still more scarce resources and to improve the quality of patient care, a number of IT vendors Endeavour in a quest to help healthcare organizations to gain control over their business processes. My research is concerned with the issues that arise when an IT-solution, based on rationalistic models of Healthcare best practices meets clinical practice and the work that needs to be done in order to align the solution to praxis or sometimes vice versa. Enterprise resource planning (ERP) is a company-wide computer software system used to manage and coordinate all the resources, information, and functions of a business from shared data stores. An ERP system has a service-oriented architecture with modular hardware and software units or "services" that communicate on a local area network. The modular design allows a business to add or reconfigure modules while preserving data integrity in one shared database that may be centralized or distributed [6] [7].

II. WHY ERP INTRODUCE IN HEALTH CARE INDUSTRY

Healthcare is yet another booming industry that enjoys unlimited growth opportunities through the globe. The obvious reason is that science and technology leads to both vices and virtues. The diseases are rapidly multiplying at a rate faster than the invention of medicines. No doubt it has become a very lucrative profession be it medical professional or paramedical professional or corporate hospitals. Some of the reasons for the increased usage and preference of ERP are as follows [9] [10]:

A. Computerization

ERP function becomes very important in this context. The attitude of personnel in hospital industry is quiet different from that of the others. When people quickly respond to computerization and automation it is practically difficult for the trend to follow suit in hospitals. The reason is doctors are keener to update about medical terminologies and surgical trends than any about any other office equipment. This has also increased the percentage of ERP use.

This could not last for a long time. The situation changed slowly. In addition to the other departments medical and paramedical professionals have taken the necessary steps to keep them abreast of the latest technology as it is directly or indirectly connected with the profession. ERP function helps to achieve this.

The concept of enterprise resource planning has made the process of segregating bills and patient records much easier. This industry also realized the need to adapt to ERP. Hence the industry was able to save more lives (but at the cost of an individual's privacy) as it enabled to access the database of patients and medical histories through the common database shared by hospitals and that too at a quicker rate.

B. Avenues for earning

The scope of medicine gets enlarged at a devastating rate. ERP architecture is an important factor in this area. What was considered as service in the **yesteryears** has become a service industry in commercial jargon. As hospitals multiply at a fast pace there is a constant urge to induce professionalism and best practices in the industry.

In this context patients demand more than the previous days. They naturally want the best return for the money. When the industry is developing at a great speed with the two elements namely business and service gets stirred in proportions. And in this situation no hospital can continue with the old age practices and technologies and still charges heavily. Therefore it becomes imperative for hospitals to adopt the latest trends in terms of technology and acumen in order to retain customer loyalty. There is no need to advocate hospitals to get the latest medical equipments or follow the medical practices. It happens naturally otherwise their survival will become a

120

question mark. However after repeated experiences and acid tests hospitals have become prune to investing in latest technology like ERP and at the same time justify their cost.

C. Pressure from patients

Hospitals were not few and far between in the olden days. Hence the patients yielded to the demands of hospitals every time even if it meant lesser quality of treatment for the charges because they had no other choice. This is evident from ERP systems definition.

The rate of hospital expansion is alarming. They have mushroomed every where. Patients are no more at the mercy of doctors but it is vice versa. When a patient is not satisfied he will not visit the hospital anymore. ERP systems definition will prove this better.

When it comes to treatment there should not be more than minor discrepancies among hospitals that are off at the same standards. One parameter for measuring the quality of treatment is the technology involved in offering it. This technology not only creates a level of comfort but also helps the patient to be confident that he is receiving the best treatment.

Consumers themselves demand tools like ERP and strongly rate it when it comes to the question of embarking with the hospital environment. There can be no wonder in saying that they have proved to be an important factor for influencing the hospitals to go for ERP. ERP review software helps them to decide the appropriate software.

D. Reduces operational costs

Enterprise resource planning helps to bring down the cost of operations. Based on the requirements the hospitals can go for best of breed or enterprise applications. This is an important step that helps hospitals to decide the appropriate software.

Hospitals can reduce their overheads through ERP as it helps to integrate all functions namely accounts , finance , human resources and bring them systems under one common database on the basis of ERP architecture.

III. ERP IN THE HEALTHCARE

ERP present an organization with operational and technical advantages, as well as a set of tangible and intangible financial benefits. In addition, ERP can provide number of benefits to health care organization. Human recourse can benefit from an ERP implementation due to centralized scheduling. In addition, for billing, laboratory, pharmacy and patient records can help in the anticipation of internal workflow.

The planning process of ERP in hospital environments:

ERP Administrator Plus integrated new generation hospital management software which converges latest technology and your administrative process to manage work process within the hospital. This is designed for multi-specialty hospitals, to cover a wide range of Hospital administration and management processes. It is an integrated client server application which uses Microsoft technologies as Front End and Flexible back end (like Oracle, SQL, etc).

Objective:

To provide an integrated Solution for the Hospital, which helps in Efficient Management of the Hospital? [11]

- Enhance Patient Care
- Improve work efficiency
- Improve Fiscal Control
- Eliminate the chances of any Pilferage

- Enable the Growth of the Hospital ERP Administrator
- Administrator Plus streamlines and integrate the operation processes and information flow in the hospital to synergize the resources namely men, material, money and equipments through information.
- This ERP facilitates hospital-wide Integrated Information System covering all functional areas like out & in Patients Billing & Management, Patient Beds, Visiting Consultants, Medical Stores, Pathology Laboratories, Radiology Laboratories, Imaging, Pharmacy, Manpower Management, Kitchen and Laundry Services etc.
- It performs core hospital activities and increases customer service thereby augmenting the overall Hospital Image. ERP bridges the information gap across the hospital.
- Administrator Plus eliminates the most of the business problems like Material shortages, Productivity enhancements, Customer service, Cash Management, Inventory problems, Quality problems, Prompt delivery, Pilferage, TPA Billing etc.

Key Features

- Administrator plus (AP) is a fully-integrated, Hospital Information System Solution.
- AP gives a total integration of order entry systems, administrative system, and departmental subsystems within a hospital.
- AP allows for scalability, reliability, efficient data processing, quick decision making, reduced maintenance and overheads.
- The data is stored in a single database providing real time data across applications throughout the hospital. Since all the data is stored centrally, it can be viewed simultaneously from multiple terminals giving all departments' access to
 - timely, up-to-date patient information.
- Administrator Plus offers a foolproof data security without user intervention to archive data.
- AP is a comprehensive information system dealing with all aspects of information processing in a hospital. This encompasses human (and paper-based) information processing as well as data processing machines.
- As an area of Medical Informatics the aim of Admin Plus is to achieve the best possible support of patient care and administration by electronic data processing.

Modules

As fig 1 ERP AP devised the following modules according to the requirements of a Multi Specialty Hospital and they integrate the various departments into a comprehensive system.



Figure 1: ERP Modules

The various modules of ERP *Administrator Plus* are [8]:

a. Reception Management

Reception is the first point of interaction for anybody coming to the Hospital. It has all the information of the patients, doctors, departments and activities of the Hospital. All enquiries and appointments are scheduled through this module. All information available here are in real time and any enquiry about the patient status, Room Status, Doctors availability or tariff's for various services is on actual status since the data is constantly updated.

This module comprises of the following components

- Patient Enquiry: This will provide information of any patient like: Patient status, Name, Address or any other demographic detail.
- Consultants Enquiry: Any Information regarding a visiting Consultant can be obtained like consultant's availability, days & time of availability, Department, specialization or any other.
- Tariff Enquiry: This option can be used to enquire about the tariffs of the hospital. The tariffs are classified into departments. Enquiry about any service provided in the Hospital can be made.
- Appointments Scheduling: This option allocates the slots for various consultants. Any appointment can be booked either by phone or visit, enquired and cancelled.
- In-patient Enquiry: Any enquiry can be made for any indoor patient in the Hospital. The enquiry can be made as per the Name, Address, Department, Bed, Ward, Patient Registration Number, etc.

b. Patient Registration

Every patient who visits the hospital has to get registered prior to getting any consultation, treatment or investigations done. Registration of patients involves accepting certain general and demographic information about the patient. The patient is allocated a unique Registration number and a Patient Identification number. The Patient ID will remain same for his all subsequent visits to the hospital whereas he will be allocated a new registration number on every visit. The consultation charges (if applicable) can also be collected for the OPD patients during registration and a receipt will be generated.

c. Out Patient Management

After registration an OPD Card is printed for the OPD patients, which list all his registration information. This card is used for the prescription writing by the consultant. An Admission form is printed with all the registration details for Indoor patients, which serves as the cover page of the patient

After the registration the patient comes to the consultation chamber, where the consultant records his history, diagnose and prescribe medicines & investigation.

The Consultant note down the following details on Patients OPD Card like Complaints, History, Diagnosis, Investigation, Medicines, Advice and Next Visit

This information is then entered into the patient data by the consultant or the operator at the OPD Counter. It serves the purpose of tracing patient's visits history and also as a feedback for research & analysis. The prescription can also be scanned and saved. The scanned data can be entered later into various fields by the operator.

d. OPD Billing

For billing of any OPD service like Pathology Tests, or any imaging investigation, the patient moves to OPD billing 122 counter. Here the services are charged as per the rates already

defined for various categories/ penal/ time etc to the patient with his Patient ID. The Payment is collected for the service provided and a receipt is generated.

This module works as an interface with the diagnostic modules. All services will be automatically entered into the respective modules wherever required like lab & Imaging reporting.

e. Investigation Reporting

In the routine functioning of a hospital, various types of investigations are carried out. Carrying out number of tests and making the results available promptly is very crucial for assessing the patient's medical status and deciding on the further course of action.

Investigation requisition can be auto-generated (through OPD billing or IPD) or can also be generated here, depending on the system followed in the hospital. The tests parameters are pre defined with the interpretations & formulae wherever applicable. The test results are entered into the software manually or with equipment integration and a descriptive smart report is printed after verification and validation.

f. Indoor Patients Management

The Indoor patient module commences when the patient is being registered and allotted a bed in the ward. It deals with the complete treatment and services provided to the patient during his stay in the hospital.

This module works at the nursing station. During his stay in the hospital, every patient is provided various services in terms of consultant's visits, investigations, procedures, medicines & consumable, room services, diet, etc. All these services are entered online to the patient record through nursing station. It also interacts with the Investigation module, Store, Pharmacy and sends the requisitions to these departments. This data serves as major input for the IPD billing.

IPD Billing

Indoor billing module has a supervisory role. The entries for billing are automatically transferred to the patient bill by the respective departments, which provide the service. The services are charged as per the category/panel/package applicable.

Here the bill is compiled and the payment collected from time to time. Provisional and Final bills are generated which provides complete information about the Services availed, its Charges, Advance collected, appropriate Receipts, Refunds, Credit notes, Concession allowed, etc.

Optional Modules

Hospital Store

This module deals with the inventory of all Hospital Equipments, Materials, Consumables, and Medicines, Implants & Asset items in different departments of the hospital along with their purchase and supplier details. Requisitions for different items/equipment are sent to this store from different departments and accordingly the Central Store issues items/equipment to various departments and generate purchase orders for purchases. This also maintains records of purchases, stock, and supplier item/equipment/material master tables.

The Store module ensures that there is a round the clock availability of a sufficient quantity of drugs and consumable material for the patients in a mode that neither hinders efficient clinical work, nor it becomes a threat to the survival of the Store.

Pharmacy

The Pharmacy Module deals with the Retail Sale of medicines to OPD patients and issue of medicines to the Inpatients in the hospital. Its function includes, online drug prescription, inventory management and billing of drugs, consumables and sutures. This module is closely linked to the Billing Module and In-patient Module. All the drugs required by the patient can be indented from the various sub stores.

Financial Accounting

A Financial accounting module is linked with hospital billing module. You get online accounting of all revenue generated along with expenses incurred. There is no need to enter the revenue entries as they are already fetched from the billing module. All relevant information for the staff salary/wages, consultant share, etc is available.

Payroll

This module Keeps track of all staff member's attendance; there leave record and deductions. Generate salary slip and other related reports.

• Canteen and Diet

This module deals with the billing of all canteen items and issue to Inpatients or staff. This will work as OPD billing. The rates of all items will be defined as per the category (general/Inpatient/staff) and the items issued to them are billed respectively. Inpatients bill can be reflected in their hospital Bill.

MRD Management

Patient's Medical record data is critical for the analysis and research purposes. This data includes patient history, observation, diagnosis and therapeutic conclusions along with the tests details and serves as a source of information for any analysis and research. The purpose for this module is to utilize the patient's medical information and use it for analysis thereby improving patient care [8].

• Online Reporting

This module transfers all the diagnostic reports performed in the day to the hospital website where the patient can view or download the report using the password. This has special significance for telemedicine.

• TPA Management

Keep track of all patients from a particular penal, their authorization amount and payment status with full bill details. Special rates for any particular penal will be automatically calculated.

Imaging Records

Image capture and analysis are vital to the use of Nuclear, CT scan, Laparoscope's, Endoscopes, Colposcopes, C-Arm & Ultrasound Systems in medicine as shown in fig 2 & 3.

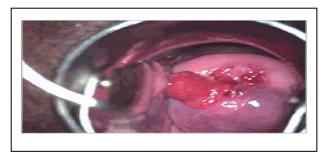


Figure 2: Image Record 1

print quality from digital medical modalities with both color and grayscale outputs with normal DeskJet or Laser printers. With our software's outstanding, High Quality Images, Superb Color Reproduction, Live Recording & Web Based utility to take 2nd opinion. Thus it can be used for any medical / scientific application or radiology solutions with digital interface and for archiving or educational purposes.

Main Features

- Can capture images indifferent formats.
- Recording facility, to record procedures or undergoing operations.
- Procedure/Operations done Live can be given on CD.
- Workstation can be created, at one point, for different modalities.
- Data base to re-analyze and to keep patient records updated.
- Can create Editing, filtering of colors(R, G, B) on the images.
- Deferent report formats / Default formats for each modality.
- Comparison mode is available.
- Exclusive help menu, with brief knowledge about the equipment & S/W.
- with voice recording.
- Clipping, Splitting and Merging of different clips (Video).
- Test overlay on a recorded video to mark abnormalities to built presentations when you need it most.



Figure 3: Image Record 2

• Special Features

- Software can be operated by Remote Key.
- Software can be operated by Foot Switch.
- Can E-mail, Report with images to take 2nd opinion.

OT Management

Detailed breakup of operation charges, OT consumables, anesthetist charges, etc from a separate module. OT scheduling for the patients is also done.

• Online MIS

This is a unique tool in the hands of top management. This module shows a screen where all current activities in terms of revenue are displayed and this screen is refreshed automatically after every 10 seconds. Simply saying, by looking at this screen, top management can find out the collection of various departments in OPD & Indoor, Outstanding of General and panel patients, No. Of admissions, Discharge of the day and so on.

IV. CONCLUSION

ERP systems have become vital strategic tools in today's competitive business environment. The paper helps to demonstrate how ERP is useful in healthcare system, what are

Digital Imaging Software dedicated for each Modality with 123 the benefits of this system, how ERP works when implemented Auto report generation utility is designed to achieve optimum

(IJCSIS) International Journal of Computer Science and Information Security, Vol. 8, No. 8, November 2010

in hospitals so that more hospitals can adopt this technology for there betterment and advancement and also work with it.

Ultimate Aim – Better Patient Care with Efficiency through ERP implementation.

REFERENCES

- [1] Why ERP? A prime on SAP Implimentation by F. Robert Jacobs, David Clay Why bark, and D. Clay Why bark
- [2] Open Source ERP by Redhuan D. Oon
- [3] ERP: Making It Happen: The Implimentation Guide to Success with Enterprise resource planning by Thomas F. wallace and Michael H. Karemzar (Hardcover - Jul 27, 2001)
- [4] What is ERP? http://www.tech-faq.com/erp.shtml
- [5] www.acsonnet.com
- [6] http://searchsap.techtarget.com/definition/ERP
- [7] http://www.mariosalexandrou.com/definition/erp.asp
- [8] ERP Solutions for Hospital Management Administrator Plus(An Integrated Hospital Management Software) http://www.acsonnet.com/index.htm
- [9] A White Paper on Healthcare ERP implementation services by Infosis 2009.
- [10] A White Paper on ERP Popularity and Need in Heath Care Industry By: Nick Mutt 2010
- [11] Critical Success Factors in Enterprise Resource Planning Implementation: A Case Study in Saudi Arabia Hospital page no.3. www.brunel.ac.uk/329/.../KhaledAlfawazpaper32.pdf.
- [12] Asian Journal of Management Research Vol., 1, No. 1, 2010 titled "A Study of Patients' Expectation and Satisfaction in Dindigul Hospitals".
- [13] Profile Soft Imaging & Medical Solutions India Pvt. Ltd.

AUTHORS PROFILE

Ms. Kirti Pancholi Lecturer Computer Application, Acropolis Institute of Pharmacutical Education and Research Indore, MP, India



Kirti Pancholi received her MCM. Degree in Computer Management from DAVV University Indore in 2005. She is LecturerComputer Applications at Acropolis Institute of Technology. Her research areas are ERP technologies, E-governance, E-Commerce, Cloud computing.

Dr. Durgesh Kumar Mishra

Professor (CSE) and Dean (R&D), Acropolis Institute of Technology and Research, Indore, MP, India,

Ph - +91 9826047547, +91-731-4730038

Email: durgeshmishra@ieee.org

Chairman IEEE Computer Society, Bombay Chapter Vice Chairman IEEE MP Subsection



Biography: Dr. Durgesh Kumar Mishra has received M.Tech. degree in Computer Science from DAVV, Indore in 1994 and PhD degree in Computer 124 Engineering in 2008. Presently he is working as Professor (CSE) and Dean

(R&D) in Acropolis Institute of Technology and Research, Indore, MP, India. He is having around 21 Yrs of teaching experience and more than 7 Yrs of research experience. He has completed his research work with Dr. M. Chandwani, Director, IET-DAVV Indore, MP, India in Secure Multi- Party Computation. He has published more than 60 papers in refereed International/National Journal and Conference including IEEE, ACM etc. He is a Senior Member of IEEE, Chairman of IEEE Computer Society, Bombay Chapter, India. Dr. Mishra has delivered his tutorials in IEEE International conferences in India as well as other countries also. He is also the programme committee member of several International conferences. He visited and delivered his invited talk in Taiwan, Bangladesh, Nepal, Malaysia, Bali-Indonesia, Singapore, Sri Lanka, USA and UK etc in Secure Multi-Party Computation of Information Security. He is an author of one book also. He is also the reviewer of tree International Journal of Information Security. He is a Chief Editor of Journal of Technology and Engineering Sciences. He has been a consultant to industries and Government organization like Sale tax and Labor Department of Government of Madhya Pradesh, India.

Fuzzy expert system for evaluation of students and online exams

¹Mohammed E. Abd-Alazeem

Computer science department, Faculty of computers and information, Mansoura Egypt,

Abstract- In this paper we will introduce an expert system for evaluation of online exam. We use fuzzy system for classifying students based on their usage data and the final marks obtained in their respective courses. We have used real data from nine Moodle courses with Mansoura University Pharmacy students and apply techniques on two hundred students. This expert system will be able to facilitate education and play the role to play the role of virtual intelligent teacher referring to student capabilities by following the feedback mechanisms and will evaluate the online exams and questions to measure the difficulty level of exams.

The main components of this expert system are Inference Engine, Knowledge Acquisition Facility and Knowledge-base that construct back-end of the system. We realize the model by a fuzzy rule-based expert system with its inference engine that uses various inference methods for education.

Keywords: Fuzzy rule base, Knowledge base, Inference engine

I. INTRODUCTION

Modern information management systems enable the recording and the management of data using sophisticated data models and a rich set of management tools. In the context of educational systems, the information typically includes details about learning material, the tasks and the objectives, the course information, the contact information, the teacher and the student profiles, and the information related to student assignments, the tests, the grades, and other records[1].

In this paper, we seek means to model the imprecision of information and simplify the access to information systems, in terms of fuzzy modeling. The paper is organized as follows. Section 2 presents the knowledge base of expert system for student grades assessment Section 3 discuss the components of fuzzy rule based controller. Section 4

² Sherief I. Barakat.

Information system department, Faculty of computers and information, Mansoura ,Egypt,

describes the results of the assessment students and online exams. Section 5 shows the feedback of evaluating students and online exams. Section 6 concludes the work and indicates future research directions.

II. Knowledgebase expert system

All needed data is acquired from a teacher and stored in a "Knowledge Based" which should be able to face student training and their up growing problem. [2] We use knowledge base expert system as follow to model knowledge domain "fig.1"

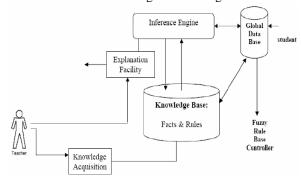


Figure 1. Knowledge base expert system

Global Data Base consists of student state variables, teaching state variables and exam state variables

- Student state variable (xi):
- x1-Exercise grade for every student
- x2-Exam grade for every student
- x3- Interesting course
- x4- Student level
- Teaching state variable (yi):
- y1- Difficulty level exam
- y2-Teaching content
- y3-Teaching method
- y4-Teaching schedule
- y5- Degree of course usages
- y6- Degree of creating motivation.

• Exam state variable (zi)

z1- Exam average grade

z2- Exam level

Fuzzy logic was primarily designed to represent and reason with some particular form of knowledge. Fuzzy logic is powerful problem solving methodology with a myriad of applications in embedded control and information processing.

Fuzzy systems are mathematically based systems that enable computers to deal with

imprecise, ambiguous, or uncertain information and situations.

Fuzzy set theory was proposed in 1965 by Zadeh to help computers reason with uncertain and ambiguous information. Zadeh proposed fuzzy technology as a means to model the uncertainty of natural language [3]. reasoned that many difficult problems can be expressed much more easily in terms of linguistic variables. Linguistic variables are words and attributes which are used to describe certain aspects of the real world. One important feature of linguistic variables is the notion of their utility as an expression of data compression. Zadeh describes this compression granulation. He argues that this is important because it is more general than use of discrete values. This point means that an agent using linguistic variables may be able to deal with more continuous and descriptions of reality and problem spaces. Our approach is to design a fuzzy rule base system to control training process.

III. FUZZY RULE BASED

This system is designed for evaluating and teaching the students so that the resulting control system will reliably and safely achieve high performance operation.

A block diagram of fuzzy system is shown in "Fig.2" Basically in fuzzy control system, there are four major stages to accomplish the control process: [4]

- Fuzzy input and output variables & their fuzzy value
- Fuzzy rule base
- Fuzzy inference engine
- Fuzzification and defuzzification modules

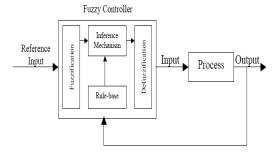


Figure 2. Fuzzy System

A. Fuzzy Inference Process

A fuzzy system works similar to a conventional system: it accepts an input value, performs some calculations, and generates an output value. This process is called the Fuzzy Inference Process and works in three steps illustrated in "Fig.3" [5]:

- Fuzzification where a crisp input is translated into a fuzzy value.
- Rule Evaluation, where the fuzzy output truth values are computed, and
- Defuzzification where the fuzzy output is translated to a crisp value.

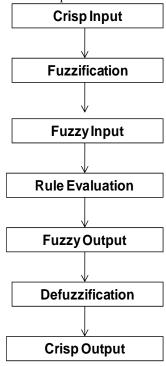


Figure 3. Fuzzy System Process

B. Fuzzification

Fuzzification where a crisp input is translated into a fuzzy value.

The membership functions defined on the input variables are applied to their actual values to determining the degree of truth.

For example for the fuzzification crisp inputs, x1 and y1 and determine the degree to which these inputs belong to each of the appropriate fuzzy sets (Figure 3).

At first it gets inputs and then fuzzifies them. After fuzzification, make decision through fuzzy inference engine according to fuzzy rule based system.

C. The Fuzzy Inference engine fuzzy rule based:

This is an interface for fuzzifying the user-requested parameters of the test items. The fuzzified parameters, along with a set of fuzzy rules, are then sent to an expert system to perform the inference process.

D. Defuzzification

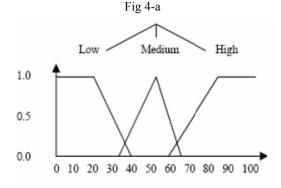
Defuzzification is a process of converting output fuzzy variable into a unique number.

Defuzzification process has the capability to reduce a fuzzy set into a crisp single-valued quantity or into a crisp set; to convert a fuzzy matrix into a crisp matrix; or to convert a fuzzy number into a crisp number. [6]

IV. EXPRIMENTS AND RESULTS

A. Evaluating an online exam

We consider two fuzzy input variables as exam average grade z1 (Fig4.a) and difficulty level of exam y1 (Fig4.b). And the output will be the exam level (z2). Membership function of z1, y1 and z2 should be as follows $(0 \le \mu \le 1)$.[7]



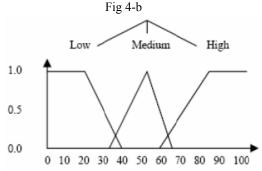


Figure 4. Membership function of exam average grade (z1) and difficulty level (y1)

Fuzzy rule base for evaluating exam is designed as follows:

R1: If z1 is high and y1 is high Then z2 is easy R2: If z1 is high and y1 is medium Then z2 is moderate

R3: If z1 is high and y1 is low Then z2 is moderate R4: If z1 is medium and y1 is high Then z2 is easy

R5: If z1 is medium and y1 is medium Then z2 is moderate

R6: If z1 is medium and y1 is low Then z2 is difficult

R7: If z1 is low and y1 is high Then z2 is moderate R8: If z1 is low and y1 is medium Then z2 is difficult

R9: If z1 is low and y1 is low Then z2 is difficult

TABLE1. Fuzzy Rules for exam evaluation

	Low	Medium	High
Low	Difficult	Difficult	Moderate
Medium	Difficult	Moderate	Easy
high	Moderate	Moderate	Easy

For Exam "Cyptology" course:

1- Exam average grade for all students (exam grade) is calculated as:

(Sum of students grades)/no of students

Exam average grade = 66%

2- Exam difficulty level

Following formula for calculating the Exam Difficulty: [8]

$$d = \frac{p}{n}$$

Where α denoted the item difficulty, β denoted the number of examinees that answered the item correctly, and β denoted the total number of examinees. Exam difficulty level = 62%

First we can apply fuzzification where a crisp input is translated into a fuzzy value,

By applying Triangle Membership Function for "Fig 4.a"

$$\mu A(z1) = 60\%$$
 $\mu A(y1) = 20\%$

$$\mu B(z1) = 20\%$$
 $\mu B(y1) = 80\%$

By applying inference mechanism

if
$$\mu A(x1) = 60\%$$
 and $\mu A(y1) = 20\%$ then

$$\mu A(z2) = 20\%$$

if
$$\mu A(z1) = 60\%$$
 and $\mu B(y1) = 80\%$ then

$$\mu A(z2) = 60\%$$

if
$$\mu B(x1) = 40\%$$
 and $\mu A(y1) = 20\%$ then

$$\mu A(z2) = 20\%$$

if
$$\mu B(z1) = 40\%$$
 and $\mu B(y1) = 80\%$ then

$$\mu A(z2) = 40\%$$

Second: Rule Evaluation, where the fuzzy output truth values are computed.

According to fuzzy based rule, we find

$$\Delta z = \mu A(z^2) = 20\%$$
 "Easy" Rule 1

 $\Delta z = \mu B(z^2) = 60\%$ "Moderate" Rule 2

Third: We will apply defuzzification where the fuzzy output is translated to a crisp value[9] shown in (figure 5).

The center of gravity is calculated as follow:

$$COG = \frac{\int_a^b \mu A(x)x dx}{\int_a^b \mu A(x) dx} = \frac{0.6*60+0.2*70}{0.2+0.6} = 62.5$$

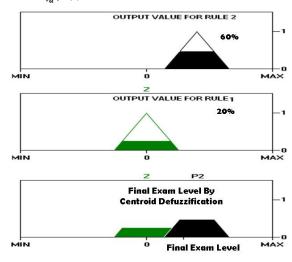


Figure 5. Defuzzification Result for Exam Evaluation

Then the exam level is "Moderate".

B. Evaluating students

We consider two fuzzy input variables as exam grade x3 (figure 6.a) and difficulty level of exam y1 (figure 6.b) and the output will be the student level (x4). Membership function of x3, y1 and x4 should be as follows $(0 \le \mu \le 1)$.

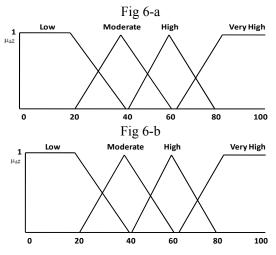


Figure 6. Membership function of student grade (x1) and difficulty level (y1)

Fuzzy rule base for evaluating student is designed as follows:

R1: If x1 is low and y1 is low Then x4 is fail

R2: If x1 is low and y1 is medium Then x4 is fail

R3: If x1 is low and y1 is high Then x4 is pass

R4: If x1 is low and y1 is very high Then x4 is pass

R5: If x1 is medium and y1 is low Then x4 is fail

R6: If x1 is medium and y1 is medium Then x4 is pass

R7: If x1 is medium and y1 is high Then x4 is good

R8: If x1 is medium and y1 is very high Then x4 is good

R9: If x1 is high and y1 is low Then x4 is pass

R10: If x1 is high and y1 is medium Then x4 is pass

R11 If x1 is high and y1 is high Then x4 is good

R12: If x1 is high and y1 is very high Then x4 is excellent

R13: If x1 is very high and y1 is low Then x4 is pass

R14: If x1 is very high and y1 is medium Then x4 is good

R15: If x1 is very high and y1 is high Then x4 is excellent

R16: If x1 is very high and y1 is very high Then x4 is excellent

TABLE 2. Fuzzy Rules for student evaluation

	Low	Medium	High	Very High
Low	Fail	Fail	Pass	Pass
Mediu	Fail	Pass	Good	Good
m				
high	Pass	Pass	Good	Excellent
Very	Pass	Good	Excellent	Excellent
High				

For student "student ID 1008":

1- Exam grade for this student is:

Student grade = 77%

2-Exam difficulty level

Exam difficulty level = 62%

First we can apply fuzzification where a crisp input is translated into a fuzzy value,

By applying Triangle Membership Function for Fig 6-a

 $\mu A(x1) = 10\%$ $\mu A(y1) = 85\%$

 $\mu B(x1) = 90\%$ $\mu B(y1) = 15\%$

By applying inference mechanism

if $\mu A(x1) = 10\%$ and $\mu A(y1) = 85\%$ then

 $\mu A(x4) = 10\%$

if $\mu A(x1)$ = 10% and $\mu B(y1)$ = 15% then $\mu A(x4)$ = 10%

if $\mu B(x1) = 90\%$ and $\mu A(y1) = 85\%$ then $\mu A(x4) = 85\%$

if $\mu B(x1) = 90\%$ and $\mu B(y1) = 15\%$ then $\mu A(z2) = 15\%$

Second: Rule Evaluation, where the fuzzy output truth values are computed.

According to fuzzy based rule, we can use $\Delta x4 = \mu A(x4) = 10\%$ "Excellent" Rule 16 $\Delta x4 = \mu B(x4) = 85\%$ "Excellent" Rule 12

Third: We will apply defuzzification where the fuzzy output is translated to a crisp value. COG = 78

Then the student level is "Excellent".

V. FEEDBACK FOR EVALUATING SRUDENTS AND ONLINE EXAM

We can classify exams according to our expert system in to 3 levels: Easy, Moderate and Difficult.[10] Then we have exam store for Pharmacy students for Mansoura University so we can evaluate this exams and give the feedback to the instructor to be a good reference for exam evaluation, so the results is as follow in Table[3]:

TABLE3. Results for exam evaluation

Course	Level
Cartilage and Bone Online Exam	Difficult
Cytology for Clinical Pharmacy Exam	Difficult
Group I Online Exam of Immune	Easy
System	
CVS Online Exam	Moderate
Urinary, Male, and Female Online	Moderate
Exam	
Online Exam of Muscular Tissue	Easy
Modifications, Glands & CT Exam	Difficult
Med-Term Exam for Clinical	Moderate
Pharmacy	
Second Med-Term Exam for CP	Difficult

The student assessment is very important because a good assessment let the instructor to have a correct decision for student follow up.

We classify student in to 4 levels: fail, pass, good and excellent. [11] So according to "Cyptology" Course the students level are as shown in Table

TABLE 4. Results for student evaluation

Student ID	Grade	Level
1006	63%	Good
1007	87%	Excellent
1008	77%	Excellent
1009	52%	Pass
1011	38%	Fail
1014	83%	Excellent
1052	60%	Good
1062	30%	Fail

CONCOLUSION

Fuzzy expert system and fuzzy rule based is a great step forward for the adaptation of the accessible knowledge for the student according to the feedback obtain from the evaluating system.

It's also considered a good reference for instructor to evaluate the exam level and the quality assurance organization is benefit from this evaluation.

ACKNOWLEDGMENT

First of all, I thank Allah for achieving this paper and giving me the ability to finish it. Second, I would like to express my appreciation to my supervisor Dr. Sherief Barakat for his continuous support and encouragement during the research study in this thesis. He really influenced my way of thinking and developing the research ideas adopted in this thesis. I am very grateful for his effort and his highly useful advice throughout the development of this work.

REFRENCES

- [1] Henry Nasution, "Design methodology of fuzzy logic control", Journal Teknos-2k, Universitas Bung Hatta, Vol.2, No.2, December (2002).
- [2] Ishiburchi, H., Nozaki, K., and Tanaka, H. "Distributed Representation of Fuzzy Rules and Its Application to Pattern Classification. Fuzzy Sets and Systems", Vol. 52,pp. 21-32. 1992.
- [3] Zadeh, L. A. "Fuzzy sets. Information and Control", Vol. 8, pp. 338-353. 1965.
- [4] Takagi, T. and Sugeon, "Fuzzy identification of System and Its Applications to Modeling and Control", vol. 15, no. 1, 116-132, 1985.
- [5] GAO Xinbo (1) XIE Weixin(2)," Advances in theory and applications of fuzzy clustering", Institute of Electronic Engineering, China, 2000.
- [6] H.Bevrani, "Defuzzification", University of Kurdistan Department of Electrical & Computer Eng, Spring Semester, 2009.
- [7] J. Harris ,"Fuzzy Logic Applications in Engineering Science", vol 29,2003.
- [8] S. J. Osterlind, "Constructing Test Items: Multiple-choice, Constructed-response,

Performance, and Other Formats", London, United Kingdom: Springer, 1998.

- [9] Sudarshan, Pavankiran, Swetha Krishnan and G Raghurama, "Fuzzy Logic Approach for Replacement Policy in Web Caching", Indian, December 2005, ISBN: 0-9727412-1-6, pp 2308-2319.
- [10] Arriaga, F. de, Alami, M. El., & Arriaga, A, "Evaluation of Fuzzy Intelligent Learning Systems". Spain, November 2005.
- [11] Nykänen, "Inducing Fuzzy Models for or Student Classification". Educational Technology & Society, vol 2, pp 223-234, 2006.

Intelligent Controller for Networked DC Motor Control

B.Sharmila

Department of EIE, Sri Ramakrishna Engineering College Coimbatore, India

N.Devarajan

Department of EEE, Government College of Tech. Coimbatore, India

Abstract—This paper focuses on the feasibility of Neural Network controller for Networked Control Systems. The Intelligent Controllers has been developed for controlling the speed of the Networked DC Motor by exploiting the features of Neural Networks and Fuzzy Logic Controllers. The major challenges in Networked Control Systems are the network induced delays and data packet losses in the closed loop. These challenges degrade the performance and destabilize the systems. The aim of the proposed Neural Network Controller and Fuzzy Logic Controller schemes improve the performance of the networked DC motor and also compare the results with the Zeigler-Nichols tuned Proportional-Integral-Derivative Controller. The performance of the proposed controllers has been verified through simulation using MATLAB/SIMULINK package. The effective results show that the performance of networked dc motor is improved by using Intelligent Controller than the other controllers.

Keywords- Networked Control Systems (NCS); Network Challenges; Tuning; Proportional – Integral - Derivative Controllers (PID); Fuzzy Logic Controller (FLC); Artificial Neural Networks (ANN).

I. INTRODUCTION

Networked Control System is the adaptation of communication network for information exchange between controllers, sensors and actuators to realize a closed control loop. Networks reduce the complexity in wiring connections and the costs of Medias. They are easy to maintain and also enable remote data transfer and data exchanges among users. Because of these benefits, many industries and institutions has shown interest in applying different types of networks for their remote industrial control and automation. Regardless of the types of networks, the overall performance of NCS is affected by two major challenges as networked induced delay and data losses. The challenges of networked DC motor are generally controlled by Conventional Proportional - Integral -Derivative Controllers, since they are less expensive with inexpensive maintenance, designed easily, and very effective. But mathematical model of the controller and tuning of PID parameters are difficult and generally not used for non-linear systems. Hence to overcome these challenges auto-tuning and adaptive PID Controller was developed with few mathematical calculations. The Intelligent controllers as Fuzzy Logic Controller and Artificial Neural Networks were used to overcoming the challenges. Thus this paper proposes Intelligent Controller for the compensation of the challenges.

The novelty of this paper lies in comparison of the application of NARMA-L2 Controller and Mamdani Fuzzy Logic Controller with conventional PID controller for the improvement of the performance of networked control DC motor.

There are two approaches to utilize a data network as Hierarchical Structure and Direct Structure as shown in Fig. 1 and Fig. 2 respectively. In the hierarchical structure the dc motor is controlled by its remote controller at remote station whereas in direct structure the central controller is used for controlling the speed of dc motor. Since the hierarchical structure has a poor interaction between central and remote unit, direct structure is preferred.

Recently the stability analysis and control design for NCS have attracted considerable research interest [3], [4], [6] and [11]. The work of Nesic and Teel [2] presents an approach for stability analysis of NCS that decouples the scheduling protocol from properties of network free nominal closed-loop system. Nesic and Tabbara [3] extended [2] by stochastic deterministic protocols in the presence of random packet dropouts and inter transmission time and they also proposed wireless scheduling protocol for non-linear NCS in [6]. The networked predictive control scheme for forward and feedback channels having random network delay was proposed in [4], and [5] addresses the problems of how uncertain delays are smaller than one sampling period which affects the stability of the NCS and how these delays interact with maximum allowable transfer interval and the selected sampling period. Robust feedback controller design for NCS with uncertainty in the system model and the network induced delay has been addressed in [7]-[8], whereas [9] handles networked induction motor speed control by using linear matrix n equality (LMI) method. Ref. [1] measure the networked vehicle control performance using an H infinity norm with linear matrix inequalities conditions and markovian jumping parameters in communication losses. In case of time varying transmission times, model based NCSs has been proposed for stabilization problem of NCS. The stability analysis and controller synthesis problems are investigated in [11] for the NCSs with random packet losses by using H infinity control and linear matrix inequalities. A moving horizon method was developed by [12], which was applied as a quantized NCS in a practical context. Since these methods transmit data specifying only a region in which the measurements lie, it will reduce the

network stabilization of the NCS. However, this method could reduce the stability of the control system by introducing uncertainty in the control system. The issues of limited bandwidth, time delay and data dropouts was taken into consideration when NCSs controllers were designed in [12] – [14]. The networked control system performance depends on the control algorithm and the network conditions. Several network conditions such as bandwidth, end-to-end delay, and packet loss rate are major impacts on networked control systems. Depending upon the control algorithm and network conditions the overall performance of the networked system may vary and hence the stability of the system.

II. MODELLING

A networked control system can be divided into the remote unit, the central controller and the data network. Fig. 3 shows the general block diagram of the networked control system under investigation. In order to focus our discussion on the performance of networked closed loop control system with network conditions (delay, data loss), a networked dc motor control system has been illustrated.

A. Remote Unit

The Remote Unit consists of the plant (dc motor), sensor and an interfacing unit. Via the network the remote unit can send measurements like motor speed, current, temperature, and local environment information, back to the central controller. The electro-mechanical dynamics of the dc motor can be described by the loop equation as first order differential equations.

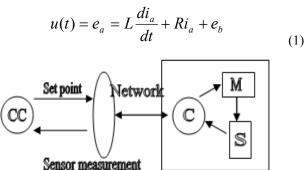
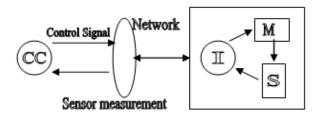


Figure 1. Hierarchical Structure.



M:Motor S:Sensor I:Interface CC:Central Controller C:Remote Contoller

Figure 2. Direct Structure.

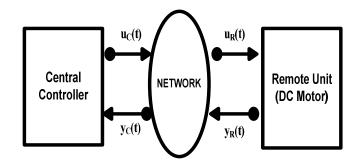


Figure 3. An overall real-time networked control system.

where $u=e_a$ is the armature winding input voltage; $e_b=K_b\omega$ is the back-electromotive-force (EMF) voltage; L is the armature winding inductance; i_a is the armature winding current; R is the armature winding resistance; K_b is the back-EMF constant and ω is the rotor angular speed. Based on Newton's law the mechanical-torque balance equation is

$$J\frac{d\omega}{dt} + B\omega + T_l = Ki_a \tag{2}$$

J is the system moment of inertia; B is the system damping coefficient; K is the torque constant and T_1 is the load torque. By letting $x_1 = i_a$ and $x_2 = \omega$, the electromechanical dynamics of the dc motor can be described by the following state-space description:

$$\dot{x}_{1}(t) = -\frac{R}{L}x_{1} - \frac{K_{b}}{L}x_{2} + \frac{1}{L}u$$

$$\dot{x}_{2}(t) = \frac{K}{J}x_{1} - \frac{B}{J}x_{2} + \frac{1}{L}T_{l}$$
(3)

The parameters of the motor Table 1 are used for determine the state space model of dc motor.

TABLE 1. DC MOTOR PARAMETERS

J	Moment of Inertia	42.6 e-6 Kg-m ²
L	Inductance	170 e-3 H
R	Resistance	4.67 Ω
В	Damping Coefficient	47.8 e-6 Nm-sec/rad
K	Torque Constant	14.7 e-3 Nm/A
Kb	Back EMF constant	14.7 e-3 Vsec/rad

B. Central Controller

The central controller will provide the control signal $u_C(t)$ to the remote systems. The central controller will monitor the network conditions of the remote unit link and provide appropriate control signals to each remote unit. Similarly the output responses are taken as feedback signal $y_R(t)$ to the central controller. The proposed Intelligent Controllers will compensate the network-induced delays, data losses and external disturbances. The data losses and disturbances occur due to missing or disturbances in input reference signal, control signal and feedback signal.

C. Data Network

There are different ways to define network conditions for point-to-point (from the central control to a specific remote unit). Two of the most popular network measures are the point-to-point network throughput and maximal delay bound of the largest data. One factor of interest is the sampling time.

To keep the illustration simple, the remote unit receives the data sent from the central controller as $u_R(t)$, which can be mathematically expressed as

$$u_R(t) = u_c(t - \tau_R) \tag{5}$$

where τ_R is the time delay to transmit the control signal $u_C(t)$ from the central controller to the remote unit. The remote unit also sends the sensors signals $y_R(t)$ of the remote system back to the central controller $y_C(t)$, and these two signals are related as

$$y_C(t) = y_R(t - \tau_C) \tag{6}$$

where τ_C is the time delay to transmit the measured signal from the remote unit to the central controller. There are also processing delays as τ_{PC} and τ_{PR} , at the central and remote unit, respectively which could be approximate small constants or even neglected because these delays are usually small compared to τ_C and τ_R .

The functions of network variables such as the network throughput, the network management/policy used, the type and number of signals to be transmitted, the network protocol used, and the controller processing time, and the network traffic congestion condition are taken as the current network conditions n(t) and let z^{-t} be a time delay operator which defines the signals as

$$u_R(t) = u_c(z^{-t_R}, n(t))$$
 (7)

$$y_c(t) = y_R(z^{-t_c}, n(t))$$
 (8)

In this paper, we have chosen sampling time as 0.5 ms and simulations are done.

III. MODELLING CONTROLLER DESIGN FOR NCS

In this session the proposed Neural Network Controller and Fuzzy Logic Controller as the central controller is described and the results are compared with the PID controller.

A. Neural Network Controller

The proposed scheme utilizes the neural-network NARMA-L2 Controller. The Neural Network Controller is designed to take the error as the input and computes the output stabilizing signal depending on the input error signal. The block diagram of Neural Network Controller for NCS is shown in Fig. 4.

The NARMA-L2 controller, a multilayer neural network has two steps involved as system identification stage and control design stage. In system identification stage, a neural network model of the plant which has to be controlled is developed and in later stage the neural network plant model has been designed to train the controller. The ANN plant specification has been shown in Table 2.

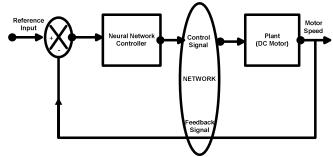


Figure 4. Neural Network Controller for NCS.

TABLE 2. ANN PLANT SPECIFICATION

No. of Inputs	3
No. of Outputs	2
No. of Hidden Layers	2
No. of Training Samples	1000
No. of Training Epochs	200

The error signals are trained for number of epochs by using the NARMA-L2 controller and the control signal are generated for any challenges in the network.

B. Fuzzy Logic controller

In general, fuzzy logic control is used for the control of a plant where the plant modeling is difficult. For such systems that are difficult to model, fuzzy logic controller has been successful by Mamdani. The basic principle of fuzzy logic lies in the definition of a set where any element can belong to a set with a certain degree of membership. Using this idea, the knowledge of an expert can be expressed in a relatively simple form and the inference for given inputs can be implemented very efficiently. Due to these advantages, fuzzy logic control is an attractive method for NCS whose modeling is very difficult because of the stochastic and discrete nature of the network. Fig. 5 shows the structure of FLC for a single input single output plant. In Fig. 5 r(t) is the reference input, y(t) is the plant output, e(t) is the error signal between the reference input and plant output and $u_C(t)$ is the control signal.

The FLC consists of three parts as 1) Fuzzifier that converts the error signal into linguistic values, 2) Inference engine that creates the fuzzy output using fuzzy control rules generated from expert experience and 3) Defuzzifier that calculate the control input to the plant from the inferred results. The input and output signals to the FLC are error signal e(t) and control signal $u_C(t)$ respectively. In this paper, the trapezoidal fuzzy members are selected for membership functions. Three fuzzy linguistic variables, i.e., Small, Medium and Large are defined. The coefficients of the membership function depend upon the set point and are determined by several trial and error experiments with the plant without the network. In order for faster execution of the fuzzy logic controller, the Mamdani's min-max inference method and the central average defuzzifier are used.

The rules used in this paper are as If e(t) is small then $u_C(t)$ is small

If e(t) is medium then $u_C(t)$ is medium If e(t) is large then $u_C(t)$ is large

C. PID Controller

It is used to compute the control signal to the remote dc motor for step tracking, based on the monitored system signals sent from the remote unit via the network link as in Fig. 6. The Proportional-Integral-Derivative (PID) controller used is

$$U_{PID}(t) = K_{p}e(t) + K_{I} \int_{0}^{t} e(t)dt + K_{D} \frac{de(t)}{dt}$$
(9)

where K_P is the proportional gain; K_I is the integral gain; K_D is the derivative gain; r(t) is the reference signal for the system to track; y(t) is the system output; and e(t) is the error function. In our case, $y = \omega$ is the motor speed, and $U_{PID}(t)$ is the input voltage to the motor system.

The results of model system with ANN, FLC and PID Controllers for network induced delays, losses and disturbance are simulated and compared.

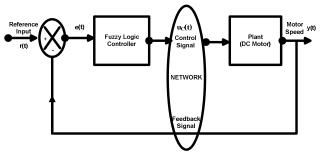


Figure 5. Fuzzy Logic Controller for NCS.

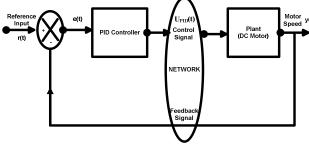


Figure 6. ZN Tuned PID Controller for NCS

IV. SIMULATION SETUP AND RESULTS

In the simulation scenario, the direct structure of the networked DC motor control system is simulated using MATLAB/ SIMULINK under fully controlled environments for Neural Network Controller, Fuzzy Logic Controller and PID Controller. Equations (3) - (4) are used as the main model, and it is controlled by the controller with the insertions of network delays according to (5) - (6). The delays are varied according to different effects of interests. The disturbance and loss of input signal, control signal and the feedback signal were made for few milliseconds at each stage and the results were studied. The system setup is illustrated in Fig.4, Fig.5 and Fig.6. Using (3)-(4) and Table 1, the state model of the dc motor is obtained. Then the results of the ANN and FLC are compared with the PID controller.

Output Responses of the system are obtained for all controllers used in this paper. Fig. 7 shows the comparison of the system performance for all controllers without delays and data losses.

Fig. 8 - 10 shows the response of the system for the controllers with different network induced delays and the comparison of these performances are tabulated in Table 3. The system responses with delay and data losses are obtained as in Fig. 11. From the simulation results as in Fig. 7 - 11, the overall system performance with Intelligent Controllers as ANN and FLC are better than the PID controller.

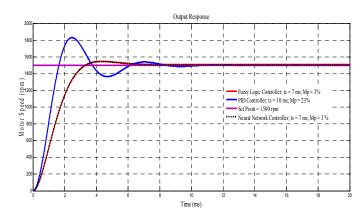


Figure 7. Comparison of System Responses for ANN, FLC and PID Controller without delay and losses.

TABLE 3. COMPARISON OF PERFORMANCE OF THE NETWORKED DC MOTOR CONTROL SYSTEM WITH DELAY IN ANN, FLC AND PID CONTROLLER. (Set point = 1500 rpm; Sampling Time = 0.5ms)

Time delay (ms)		Maximum overshoot (%)			Settling Time (ms)		
Feedforward path	Feedback path	PID	FLC	ANN	PID	FLC	ANN
0.5	1	3.3	3.3	3	30	7	7
1	1	3.3	3.3	3	40	8	7
2	2	6.6	3.3	3	62	9	8
2	3	8	3.3	3	70	9	8
3	2	9	3.3	3	75	10	9

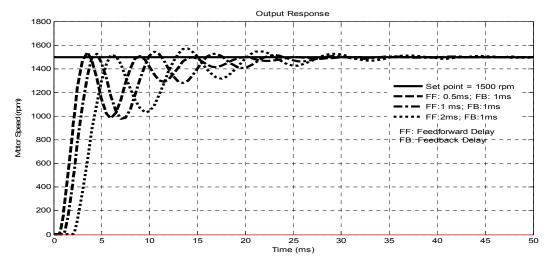


Figure 8. Response of the System using PID Controller with varying delays in forward and feedback path of NCS.

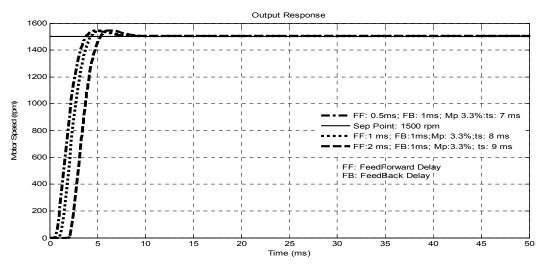


Figure 9. Response of the System using FLC with varying delays in forward and feedback path of NCS.

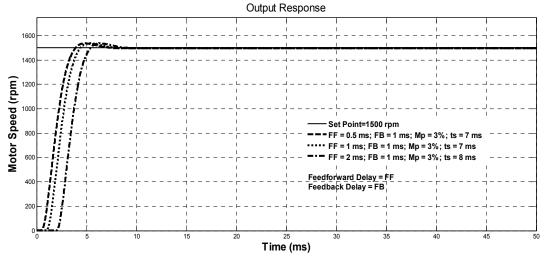


Figure 10. Output Response of the System using ANN with varying delays in forward and feedback path of NCS.

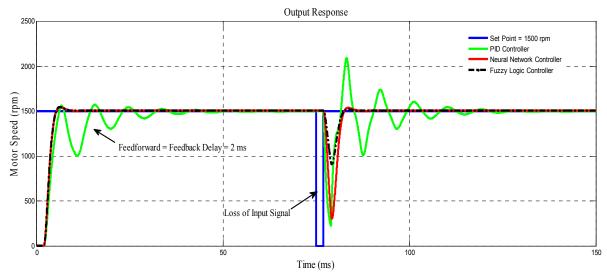


Figure 11. Comparison of system responses of ANN, FLC and PID Controllers with delay and losses.

V. CONCLUSION

Networks and their applications play a promising role for real-time high performance networked control in industrial applications. The major concerns are the network induced delays and data losses that are provided by the network which affects the performance of the networked control systems. This paper has describes and formulates the Intelligent Controllers as Neural Network Controller and Fuzzy Logic Controller in a networked DC motor control. The numerical result are obtained and compared for Neural Network Controller, Fuzzy Logic Controller and PID Controller. The effective results show that the performance of networked control DC motor is improved by using Intelligent Controller than the convention controller in all network variations and deteriorations. The analysis on using intelligent controls improves and strengthens the networked control systems concepts in the future.

REFERENCES

- [1] P.Seiler, and R. Sengupta, "An H∞ Approach to Networked Control," IEEE Trans. Autom. Control, vol. 50, pp. 356-364, March 2005..
- [2] D.Nesic, and A.R.Teel, "Input-Output stability properties of networked control systems," IEEE Trans. Autom. Control, vol. 49, pp. 1650-1667. October 2004.
- [3] M.Tabbara, and D. Nesic, "Input-Output Stability of Networked Control Systems With Stochastic Protocols and Channels," IEEE Trans. Autom. Control, vol. 53, pp. 1160-1175, June 2008.
- [4] G.P.Lin, Y. Xia, J.Chen, D.Rees, and W.Hu, "Networked Predictive Control of Systems With Random Network Delays in Both Forward and Feedback Channels," IEEE Trans. Ind. Electron., vol. 54, pp. 1282-1297, June 2007.
- [5] D.S. Kim, Y.S. Lee, W.H. Kwon, and H.S.Park, "Maximum allowable delay bounds of networked control system," Control Eng. Practice, vol. 11, pp. 1301–1313, 2003.
- [6] M.Tabbara, C. Nesic, and A.Teel, "Stability of wireless and wireline networked control systems," IEEE Trans. Autom. Control, vol. 52, pp. 1615-1630, September 2007.
- [7] D. Yue, Q. Han, and P. Chen, "State feedback controller design of networked control systems," IEEE Trans. Circuits Systems II, vol. 51, pp. 640-644, November 2004.

- [8] D.Yue, Q.Han, and J.Lam, "Network-based robust H∞ control of systems with uncertainty," Automatica, vol. 41, pp. 999-1007, June 2005.
- [9] J.Ren, C.Wen Li De, and Z. Zhao, "Linearizing Control of Induction Motor Bsed on Networked Control Systems," International Journal of Automation and Computing, vol. 6, pp. 192-197, May 2009.
- [10] L.A.Montestruque, and P. Antsaklis, "Stability of Model-Based Networked Control Systems with Time-Varying Transmission Times," IEEE Trans. Autom. Control, vol. 49, pp. 1562–1571, September 2004.
- [11] Z.Wang, F.Yang, W.C.H.Daniel and X.Liu, "Robust H

 Control for Networked Systems with Random Packet Losses," IEEE Trans. Sys. Man Cybernetics-Part B, vol. 37, pp. 916-923, August 2007.
- [12] G.C.Goodwin, H.Haimovich, D.E.Quevedo, and J.S.Welsh, "A Moving Horizon approach to networked control system design," IEEE Trans. Autom. Control, vol. 49, pp. 1427–1445, September 2004.
- [13] K.Li, and J.Baillieul, "Robust quantization for digital finite communication bandwidth (DFCB) control," IEEE Trans. Autom. Control, vol. 49, pp. 1573–1584, September 2004.
- [14] R.C.Luo, and T.M.Chen, "Development of a multibehaviour -based mobile robot for remote supervisory control through the internet," IEEE. Trans. Mechatron., vol. 5, pp. 376-385, October 2000.
- [15] J.P.Hespanha, P.Naghshtabrizi, and Y.Xu, "A survey of recent results in networked control systems," Proc. IEEE., vol. 95, pp. 138-162, January 2007
- [16] Y.Tipsuwan, and M.Y.Chow, "Control methodologies in networked control systems," Control Eng. Practice, vol. 11, pp. 1099-1111, Feburary 2003.
- [17] K.Ogata, Modern Control Engineering, Englewood Cliffs, NJ: Prentice Hall, 1990.
- [18] J.G.Ziegler, and N.B.Nichols, "Optimum settings for automatic controllers," Trans. ASME, vol. 64, pp. 759-768, November 1942.
- [19] C.C.Lee, "Fuzzy logic in control systems: fuzzy logic controller-Part I, "IEEE Trans. Syst., Man, Cybern., vol. 20, pp. 404-418, Mar/Apr 1990.
- [20] C.C.Lee, "Fuzzy logic in control systems: fuzzy logic controller-Part II," IEEE Trans. Syst., Man, Cybern., vol. 20, pp. 419-435, Mar/Apr 1990.

AUTHORS PROFILE

Dr.N.Devarajan received B.E (EEE) and M.E (Power Electronics) from GCT Coimbatore in the year 1982 and 1989 respectively. He received Ph.D in the area of control systems in the year 2000. He is currently working as Assistant Professor in the department of EEE at Government College of Technology, Coimbatore. He published 135 papers in the national and international

(IJCSIS) International Journal of Computer Science and Information Security, Vol. 8, No.8, 2010

conferences. He published 37 papers in international journals and 12 in national journal. Under his supervision currently 10 research scholars are working and 7 scholars completed their Ph.D. His areas of interests are control systems, electrical machines and power systems. He is a member of system society of India, ISTE and IE(India).

B.Sharmila completed B.E (EIE) from Tamilnadu College of Engineering, Coimbatore in the year 2000. She completed her M.E (Applied Electronics)

from Maharaja College of Engineering, Coimbatore in the year 2004. She is currently working as Senior Lecturer in the department of EIE at Sri Ramakrishna Engineering College, Coimbatore. She is a Ph.D. research scholar and published 2 papers in international journals and also presented 4 papers in national and international conference. Her areas of interests are networked control system and intelligent controllers. She is a member of IEEE and

A Novel LTCC Bandpass Filter for UWB Applications

Thirumalaivasan K and Nakkeeran R
Department of Electronics and Communication Engineering
Pondicherry Engineering College, Puducherry-605014, India

Abstract— Bandpass filter based on parallel coupled line microstrip structure is designed in low-temperature co-fired ceramic technology (LTCC) suitable for short range Ultra-Wideband (UWB) applications. Fifth order Chebyshev filter of 0.05 dB passband ripple with fractional bandwidth of 62.17% is proposed using insertion loss method. The filter demonstrates -10 dB bandwidth and linear phase response over the frequency range 3.8 GHz - 7.4 GHz. With the above functional features, the overall dimension of the filter is 33.5 mm (height) \times 1.6 mm (length) \times 1.6 mm (breadth). It is not only compact but also delivers excellent scattering parameters with the magnitude of insertion loss, $|\mathbf{S}_{21}|$ lower than -0.09 dB and return loss better than -49 dB. In the passband, the computed group delay is well within the tolerable variation of 0.1 ns.

Keywords- Ultra-wideband; bandpass filter; parallel coupled line; low-temperature co-fired ceramic; group delay

I. INTRODUCTION

UWB technology has brought out tremendously increasing research interests since the Federal Communications Commission (FCC) in USA released its unlicensed use for indoor and hand-held systems in 2002 [1]. Efforts have been made in the past eight years towards exploring various UWB components and devices. As one of the essential component blocks, the researchers are attempting to design the UWB bandpass filter (BPF) with 120% fractional bandwidth centered at 6.85 GHz. In the recent years, the market pays much attention towards miniaturization of receiver systems. Hence, researchers are working for the development of small size and cost effective filters [2]-[5].

Parallel coupled-line microstrip filters are found to be one of the most commonly used microwave filters in many practical wireless systems for several decades [6]-[8]. In addition to the planar structure and relatively wide bandwidth, the major advantage of this kind of filter is that its design procedure is quite simple. Based on insertion loss method [9], filter functions of maximally flat and Chebyshev type can be easily synthesized. Moreover, the filter performance can be improved in a straightforward manner by increasing the order of the filter. When these filters are to be realized by parallel coupled microstrip lines, one of the main limitations is the small gap size of first and last coupling stages. To increase the coupling efficiency, more fractional bandwidth and smaller

gap size are required. Obviously, shrinking the gap size is not only the way to increase the coupling of coupled lines [10].

The proposed bandpass filter in this paper is based on LTCC using parallel coupling at center and broad side coupling at ends of the proposed filter structure. The filter is designed to cover the entire UWB range. The main advantage of the multi-layered structure is to shrink the circuit size. The obtained scattering parameters of UWB bandpass filter convey an optimal performance in terms of insertion and return loss. It is distinctive in its structure and it has simple design with less number of design parameters compared to the existing filter designs in the literature [11]-[13].

The rest of the paper is organized as follows: In Section II, the UWB bandpass filter design using LTCC is presented. Simulation results and analysis are presented in Section III. Section IV concludes the paper.

II. BANDPASS FILTER DESIGN

Figure 1 shows one possible circuit arrangement for bandpass filter using parallel coupled line microstrip structure at center and broad side coupling at end of the geometry designed in LTCC for UWB range. It consists of transmission line sections having the length of half wavelength at the corresponding center frequency. Half wavelength line resonators are positioned so that adjacent resonators are parallel to each other along half of their length. This parallel arrangement gives relatively large coupling for the given spacing between the resonators, and thus, this filter structure is particularly convenient for constructing filters having larger bandwidth as compared to the other structures [14]-[17].

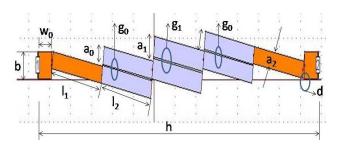


Figure 1. Geometry of the proposed UWB bandpass filter

The gap between the resonators is introducing a capacitive coupling, which can be represented by a series capacitance. The broad side coupling and existence of the substrate result tight coupling, which provides wide bandpass operation. The physical parameters of the proposed bandpass filter are optimized to the following values, l₁=5.89 mm; l₂=5.86 mm; d=0.2 mm; $g_0 = 0.08$ mm; $g_1 = 0.1$ mm; $a_1 = 0.06$ mm; $a_1 = 1.2 \text{ mm}$; $a_2 = 1.02$; $w_0 = 1.6 \text{ mm}$; b = 1.6 and h = 33.5 mmto cover the entire UWB range between 2 GHz and 9 GHz. Using this configuration, higher coupling is obtained and therefore wider bandwidth is achieved. This structure is used to generate a wide passband and expected to achieve a tight coupling, and lower insertion by reducing both strip and slot width. 3D view of a LTCC UWB bandpass filter with parallel and broadside coupling is shown in Figure 2, which consists of two layers, resonators and substrate with frames.

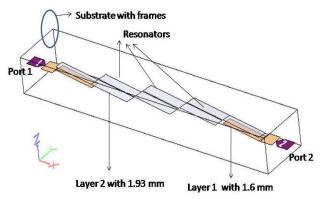


Figure 2. 3D view of proposed LTCC filter structure

III. RESULTS AND DISCUSSION

The proposed filter is designed to provide a wide passband, low insertion loss and return loss, linear phase over the passband, flat group delay and high fractional bandwidth. The simulation S parameters of the proposed UWB bandpass filter using LTCC are shown in Figure 3. It is clear from the response that the proposed filter has better insertion loss of -0.09 dB and the low return loss of about -49 dB. The -10 dB fractional bandwidth computed from the response is about 62.17 %.

For wideband applications, the examination of the flat group delay is essential and required. The simulation group delay for the proposed filter is shown in Figure 4, which exhibits a flat group delay response below 0.1 ns over the whole passband. It implies that this proposed UWB filter has a very good linearity of signal transfer and would ensure the minimum distortion to the input pulse when it is implemented in the UWB system. The response of the Figure 5 shows that the phase of S_{21} throughout the -10 dB passband between 3.8 GHz and 7.4 GHz of designed filter is acceptably linear.

In order to evaluate the performance of the proposed UWB bandpass filter, the filter is simulated through the simulation tool, IE3D [18]. The filter is designed based on LTCC substrate with two upper sheet layers, thickness of 1.6 mm and 1.93 mm with dielectric constant of 7.8 and a loss tangent of 0.002.

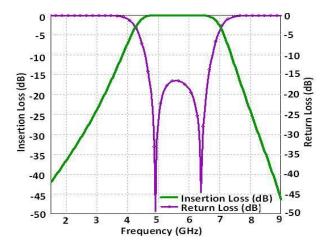


Figure 3. Simulation S- parameters

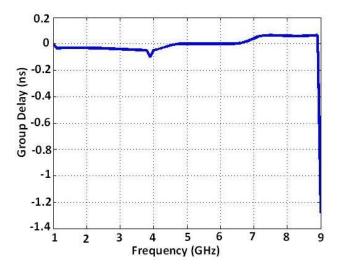


Figure 4. Simulation group delay

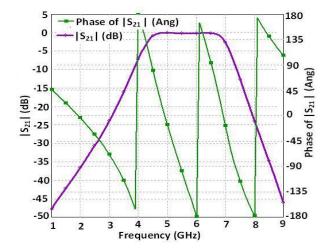


Figure 5.Simulation of phase of S₂₁

Vol. 8, No. 8, November 2010

CONCLUSION

In this letter, a bandpass filter for UWB applications based on LTCC structure is presented. The proposed filter demonstrated an excellent ultra-wide bandwidth from 3.8 GHz to 7.4 GHz. Total size of the UWB filter is 33.5 mm (height) \times 1.6 mm (length) \times 1.6 mm (breadth) and the fractional bandwidth is about 62.17 %. Simulation of bandpass filter delivers excellent scattering parameters with magnitude of insertion loss, $|S_{21}|$ lower than -0.09 dB and return loss better than -49 dB. The obtained group delay for this filter is below 0.1 ns.

REFERENCES

- [1] FCC NEWS (FCC 02-48), Feb. 14, 2002. FCC News release.
- [2] Hong, J. S. and M. J. Lancaster, Microstrip Filters for RF/Microwave Application, Wiley, New York, 2001.
- [3] C.Q.Scrantom and J.C.Lawson, "LTCC Technology where we are and where we're going-II", In IEEE MTT Int. Microwave. Symp. Dig. 1999.,pp.193-200
- [4] C.W.Tang, Harmonic Suppression LTCC Filter with the Step Impedance Quarter Wavelength Open stub", *IEEE Trans. Microw. Theory Tech.*, vol. 51, no. 10, pp. 2112–2118, Oct. 2003.
- [5] Jorge A. Ruiz Cruz, Yunchi Chang, Kawthar A. Zaki, Andrew J.Piloto and Joseph Tallo, "Ultra-Wideband LTCC Ridge Waveguide Filters,", " *IEEE Microw. Wireless Compon. Lett.*, vol. 17, no. 2, pp. 111-117, Feb. 2007.
- [6] Jen-Tsai Kuo, Wei-Hsiu Hsu, and Wei-Ting Huang," Parallel Coupled Microstrip Filters with Suppression of Harmonic Response," *IEEE Microw. Wireless Compon. Lett.*, vol. 12, no.10, Oct. 2002.
- [7] L. Zhu, S. Sun, and W. Menzel, "Ultra-wideband (UWB) bandpass filters using multiple-mode resonator," *IEEE Microw. Wireless Compon. Lett.*, vol. 15, no. 11, pp. 796-798, Nov. 2005.
- [8] Hussein Shaman, Jia-Sheng Hong," Asymmetric Parallel-Coupled Lines for Notch Implementation in UWB Filters," *IEEE Microw. Wireless Compon. Lett.*, vol. 17, no.7, July 2007.
- [9] Pozar, D. M., Microwave Engineering, Wiley, New York, 1998.
- [10] T.-N. Kuo, S.-C. Lin, and C. H. Chen, "Compact ultra-wideband bandpass filters using composite microstrip-coplanar-waveguide structure," *IEEE Trans. Microw. Theory Tech.*, vol. 54, no. 10, pp. 3772–3777, Oct. 2006.
- [11] Oudayacoumar S, Nakkeeran R and Thirumalaivasan K, "Resonator Based Compact Ultra-Wideband and Notched Wideband Filters", in Proceedings of the IEEE National Conference on Communication (NCC 2010), at IIT Chennai, January 29-30, 2010. DOI:10.1109/NCC.2010.5430168.

- [12] Thirumalaivasan K and Nakkeeran R, "Wired Ring Resonator Based Compact Ultra-Wideband Bandpass Filters Using Dual-line Coupling Structure", in the Proc. of ACEEE International Conference on Control, Communication and Power Engineering CCPE 2010, July 28-29, 2010, India. DOI: 02.CCPE.2010.1.182.
- [13] J.-T. Kuo, "Accurate quasi-TEM spectral domain analysis of single and multiple coupled microstrip lines of arbitrary metallization thickness," *IEEE Trans. Microwave Theory Tech.*, MTT-43, no.8, pp. 1881-1888, Aug. 1995.
- [14] Thirumalaivasan K, Nakkeeran R, and Oudayacoumar S, "Circular Resonator Based Compact Ultra-Wideband Bandpass and Notched Filters with rejection of 5-6 GHz band", in the *Proc. of ACEEE International Conference on Control, Communication and Power Engineering CCPE 2010*, July 28. DOI: 02.CCPE.2010.1.212.
- [15] Yue Ping Zhang and Mei Sun, "Dual Band Microstrip Bandpass Filter Using Stepped Impedance Resonators With New Coupling Schemes," *IEEE Trans. on Microwave Theory and Tech.*, vol.54, no.10, Oct 2006.
- [16] Thirumalaivasan K, Nakkeeran R, and Oudayacoumar S, "Effective Notch Ultra-Wideband Filter Using Ring Resonator for the Rejection of IEEE 802.11a", in the Proc. of IEEE International Conference on Computing, Communication and Networking ICCCNT 2010, July 29. DOI:10.1109/ICCCNT.2010.5592565.
- [17] Ravee Phromloungsri, Mitchai Chongcheawchamnan and Ian D. Robertson, "Inductively Compensated Parallel Coupled Microstrip Lines and Their Applications," *IEEE Trans. on Microwave Theory and Tech.*, vol.54, no.9, Sep 2006.
- [18] IE3D 14, Zeland Software, Ins., Fremont, USA

AUTHORS PROFILE

Mr.K.Thirumalaivasan was born in India. He received the B.Tech. degree in Electronics and Communication Engineering from Pondicherry University, Puducherry, India, and the M.E. degree in Communication Systems from College of Engineering Guindy, Anna University, Chennai, India, in 2004 and 2007 respectively. He is currently working towards the Ph.D. degree at Pondicherry Engineering College, Pondicherry. His current research interest is in the area of UWB filters and narrowband interference issues with UWB systems.

Dr. R. Nakkeeran Received BSc. Degree in Science and B.E degree in Electronics and Communication Engineering from the Madras University in 1987 and 1991 respectively and M.E degree in Electronics and Communication Engineering (diversification in Optical Communication) from the Anna University in 1995. He received Ph.D degree from Pondicherry University in 2004. Since 1991, he has been working in the teaching profession. Presently, he is Associate Professor in Pondicherry Engineering College. He is life member of IETE, ISTE, OSI and IE (I). Also he is member of OSA, SPIE and IEEE. He has published seventy five papers in National and International Conference Proceedings and Journals. He has co-authored a book, published PHI. His areas of interest are Optical Communication, Networks, Antennas, Electromagnetic Fields and Wireless Communication.

Retrieval of Bitmap Compression History

Salma Hamdy, Haytham El-Messiry, Mohamed Roushdy, Essam Kahlifa Faculty of Computer and Information Sciences Ain Shams University Cairo, Egypt

Abstract—The histogram of Discrete Cosine Transform coefficients contains information on the compression parameters for JPEGs and previously JPEG compressed bitmaps. In this paper we extend the work in [1] to identify previously compressed bitmaps and estimate the quantization table that was used for compression, from the peaks of the histogram of DCT coefficients. This can help in establishing bitmap compression history which is particularly useful in applications like image authentication, JPEG artifact removal, and JPEG recompression with less distortion. Furthermore, the estimated table calculates distortion measures to classify the bitmap as genuine or forged. The method shows good average estimation accuracy of around 92.88% against MLE and autocorrelation methods. In addition. because bitmaps do not experience data loss, detecting inconsistencies becomes easier. Detection performance resulted in an average false negative rate of 3.81% and 2.26% for two distortion measures, respectively.

Keywords: Digital image forensics; forgery detection; compression history; Quantization tables.

I. Introduction

Although JPEG images are the most widely used image format, sometimes images are saved in an uncompressed raster form (bmp, tiff), and in most situations, no knowledge of previous processing is available. Some applications are required to receive images as bitmaps with instructions for rendering at a particular size and without further information. The image may have been processed and perhaps compressed with contain severe compression artifacts. Hence, it is useful to determine the bitmap history; whether the image has ever been compressed using the JPEG standard and to know what quantization tables were used. Most of the artifact removal algorithms [2-9] require the knowledge of the quantization table to estimate the amount of distortion caused by quantization and avoid over-blurring. In other applications, knowing the quantization table can help in avoiding further distortion when recompressing the image. Some methods try to identify bitmap compression history using Maximum Likelihood Estimation (MLE) [10-11] or by modeling the distribution of quantized DCT coefficients, like the use of Benford's law [12], or modeling acquisition devices [13].

Furthermore, due to the nature of digital media and the advanced digital image processing techniques, digital images may be altered and redistributed very easily forming a rising threat in the public domain. Hence, ensuring that media content is credible and has not been altered is becoming an important issue governmental security and commercial applications. As a result, research is being conducted for developing authentication methods and tamper detection techniques. Usually JPEG compression introduces blocking artifacts and hence one of the standard passive approaches is to use inconsistencies in these blocking fingerprints as a reliable indicator of possible tampering [14]. These can also be used to determine what method of forgery was used.

In this paper we are interested in the authenticity of the image. We extend the work in [1] to bitmaps and use the proposed method for identifying previously compressed bitmaps and estimating the quantization table that was used. The estimated table is then used to determine if the mage was forged or not by calculating distortion measures.

In section 2 we study the histogram of DCT AC coefficients of bitmaps and show how it differs for previously JPEG compressed bitmaps. We then validate that without modeling rounding errors or calculating prior probabilities, quantization steps of previously compressed bitmaps can still be determined straightforward from the peaks of the approximated histograms of DCT coefficients. Results are discussed in section 3. Section 4 is for conclusions.

II. HISTOGRAM OF DCT COEFFICIENTS IN BITMAPS

We studied in [1] the histogram of quantized DCT coefficients and showed how it can be used to estimate quantization steps. Here, we study uncompressed images and validate that the approximated histogram of DCT coefficients can be used to determine compression history. Bitmap image means no data loss and hence all what is required to build an informative histogram is expected to be present in the coefficients histograms.

The first step is to decide if the test image was previously compressed because if the image was an original uncompressed there is no compression data to extract. When the image is decided to have a compression history, the next step is to estimate that history. For grayscale image, compression history mainly means its quantization table which will be the focus of this paper. For color image, this is extended to estimating color plane compression parameters that includes subsampling and associated interpolation.

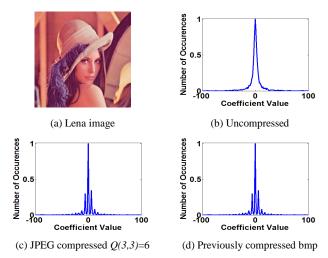


Fig. 1. Histograms of $X^*(3,3)$.

Fig. 1(b) shows the approximated histogram H^* of DCT coefficient at position (3,3) of the luminance channel of an uncompressed Lena image and the histogram of the image after being JPEG compressed with quality factor 80. It is clear that the latter contains periodic patterns that are not present in the uncompressed version. It was observed that the coefficient is very likely to have been quantized with a step of this periodic [15]. Now if that JPEG was stored in a bitmap uncompressed form, we expect the DCT coefficients to have the same behavior because nothing is lost during this format change. This is evident in Fig. 1(d) which shows an identical histogram to the one in Fig. 1(c). Hence, similar to the argument in [1], if we closely observe the histogram of $H^*(i,j)$ outside the main lobe, we notice that the maximum peak occurs at a value that is equal to the quantization step used to quantize $X_q(i,j)$. This observation applies to most low frequency AC coefficients. Fig. 2(a) and (b) show |H|, the absolute histograms of DCT coefficients for Lena of Fig. 1(a) at frequencies (3,3) and (3,4), respectively. As for high frequencies, the maximum occurred at a value matching Q(i,j)when $|X^*(i,j)| > B$, (**Fig. 2** (c) and (d)), where B is as follows:

$$\Gamma = \left| X^{*}(i,j) - X_{q}(i,j) \right| \le B(i,j)$$

$$= \sum_{u,v} 0.5 \ c(u) \ c(v) \left| \cos \frac{(2u+1)i\pi}{16} . \cos \frac{(2v+1)j\pi}{16} \right|$$
 (1)

where $X_q(i,j)$ is the quantized coefficient, and $X^*(i,j)$ is the approximated quantized coefficient, Γ is the round off error,

and
$$c(\omega) = \begin{cases} 1/\sqrt{2} & \text{for } \omega = 0\\ 1 & \text{otherwise} \end{cases}$$

See [1, 11].

Sometimes we do not have enough information to determine Q(i,j) for high frequencies (i,j). This happens when the histogram outside the main lobe decays rapidly to zero showing no periodic structure. This reflects the small or zero value of the coefficient. At such cases, it can be useful to

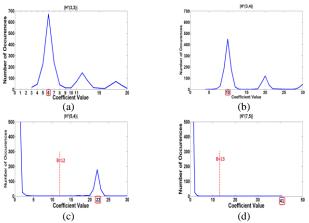


Fig. 2. (a) $|X^*(3,3)|$ where H_{max} occurs at Q(3,3)=6. (b) $|X^*(3,4)|$ where H_{max} occurs at Q(3,4)=10 (c) $|X^*(5,4)|$ where H_{max} occurs at Q(5,4)=22. (d) $|X^*(7,5)|$ where H_{max} occurs at Q(7,5)=41.

estimate as many of the low frequencies and then search through lookup tables for a matching *standard* table.

Estimating the quantization table of a bitmap can help determine part of its compression history. If all (or most of) of the low frequency steps were estimated to be ones, we can conclude that the image did not go through previous compression. High frequencies may bias because they have very low contribution and do not provide a good estimate. Moreover, this method works well also for uncompressed or lossless compressed tiff images. Fig. 3(d) shows the 96.7% correctly estimated Q table using the above method of a tiff image taken from UCID [16]. The X's mark the "undetermined" coefficients.

Now for verifying the authenticity of the image, we use the same distortion measures we used in [1]. The average distortion measure is calculated as a function of the remainders of DCT coefficients with respect to the original Q matrix:

$$B_1 = \sum_{i=1}^{8} \sum_{j=1}^{8} \operatorname{mod}(D(i, j), Q(i, j))$$
 (2)

where D(i,j) and Q(i,j) are the DCT coefficient and the corresponding quantization table entry at position (i,j), respectively. An image block having a large average distortion value indicates that it is very different from what it should be and is likely to belong to a forged image. Averaged over the entire image, this measure can be used for making a decision about authenticity of the image.

In addition, the JPEG 8×8 "blocking effect" is somehow still present in the uncompressed version and hence blocking artifact measure, BAM [14], can be used to give an estimate of the distortion of the image. It is computed from the Q table as:

$$B_{2}(n) = \sum_{i=1}^{8} \sum_{j=1}^{8} \left| D(i,j) - Q(i,j) \, round\left(\frac{D(i,j)}{Q(i,j)}\right) \right| \tag{3}$$

where B(n) is the estimated blocking artifact for the n^{th} block.



(a) Test image

	3	4	4	6	10	16	20	24
	5	5	6	8	10	23	24	22
	6	5	6	10	16	23	28	22
	6	7	9	12	20	35	32	25
	7	9	15	22	27	44	41	31
I	10	14	22	26	32	42	45	37
	20	26	31	35	41	48	47	Х
	29	37	38	39	45	40	Х	Х

(c) Estimated Q for previously compressed version with QF = 80.

Fig. 3. Estimating *Q* table for original and previously compressed tif image.

1 1 1 1 1 10 10 10 1 10 10 10 1 1 1 10 10 10 1 1 14 12 12 12 1 12 13 11 11 1 11 12 11 13 12 12 12

(b) Estimated O for uncompressed version (most low frequencies are ones).

	3	0	0	0	0	0	0
Ī	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0	0	0	1	Х
Ī	0	0	0	0	0	Х	Х

(d) Difference between (c) and original table for QF=80.

III. EXPERIMENTAL RESUTLS AND DISCUSSION

A. Estimation Accuracy

Our testing image set consisted of 550 images collected from different sources (more than five camera models), in addition to some from the public domain Uncompressed Color Image Database (UCID), which provides a benchmark for image processing analysis [16]. Each of these images was compressed with different quality factors, [60, 70, 80, and 90]. Again, each of these was uncompressed and resaved as bitmap. This yielded $550\times4 = 2,200$ untouched images. For each quality factor group, an image's histogram of DCT coefficients at one certain frequency was generated and used to determine the corresponding quantization step at that frequency according to section 2. This was repeated for all the 64 histograms of DCT coefficients. The resulting quantization table was compared to the quality factor's known table and the percentage of correctly estimated coefficients was recorded. Also, the estimated table was used in equations (2) and (3) to determine the image's average distortion and blocking artifact measures, respectively. These values were recorded and used later to set a threshold value for distinguishing forgeries from untouched images.

Table 1 shows the accuracy of estimating all 64 entries using the proposed method for each quality factor averaged over the whole set. It exhibits a similar behavior to JPEG images; as quality factor increases, estimation accuracy increases steadily with an expected drop for quality factors higher than 90 as the periodic structure becomes less prominent and the bumps are no longer separate enough . Overall, we can see that the estimation accuracy is higher than

TABLE I. PERCENTAGE OF CORRECTLY ESTIMATED COEFFICIENTS FOR SEVERLA QFS

	QF	60	70	80	90
BMP		82.07%	84.80%	87.44%	89.44%
JPEG[1]		72.03%	76.99%	82.36%	88.26%

that of JPEG images [1]. We anticipate that because lossy compression tends to lessen available data to make a better estimate. Average estimation time for all 64 entries of images of size 640×480 for different QFs was 52.7 seconds.

Estimating Q using MLE methods [10-11] is based on searching for all possible Q(i,j) for each DCT coefficient over the whole image which can be computationally exhaustive for large size files. Another method [12] proposed a logarithmic law and argued that the distribution of the first digit of DCT coefficients follows that generalized Benford's law. The method is based on re-compressing the test image with several quality factors and fitting the distribution of DCT coefficients of each version to the proposed law. The QF of the version having the least fitting artifact is chosen and its corresponding Q table is the desired one. Of course the above methods can only estimate standard compression tables. Although it may be accurate, it is time consuming. Plus it fails when the recompression quantization step is an integer multiple of the original compression step size. Another method [17] tends to calculate the autocorrelation function of the histogram of DCT coefficients. The displacement corresponding to the peak closest to the peak at zero is the value of Q(i,j) given that the peak is higher than the mean value of the autocorrelation function. The method eventually uses a hybrid approach; the low frequency coefficients are determined directly from the autocorrelation function, while the higher-frequency ones are estimated by matching the estimated part to standard JPEG tables scaled by a factor of s, which is determined from the known coefficients.

Table 2 shows the estimation accuracy while Table 3 shows estimation time, for the different mentioned methods against ours. Note that accuracy was calculated for directly estimating only the first nine AC coefficients without matching. This is due to the methods failing to estimate high frequency coefficients as most of them are quantized to zero. On the other hand, the listed time is for estimating the nine coefficients and then retrieving the whole matching table from JPEG standard lookup tables. Maximum peak is faster than

TABLE II. ESTIMATION ACCURCAY FOR THE FIRSY 3×3 AC COEFFICIENTS FOR SEVERAL OFS

QF Method	50	60	70	80	90	100	Avg. Acc.
MLE	75.31	83.10	90.31	96.34	93.83	59.5	83.06
Benford	99.08	87.59	80.82	93.81	59.47	31.53	75.38
Auto.	48.94	50.37	63.71	81.43	65.37	57.50	61.22
Max.Peak	97.93	97.07	99.01	97.67	89.57	76.04	92.88

TABLE III. ESTIMATION TIME IN SECONDS FOR THE FIRSY 3×3 AC COEFFICIENTS FOR SEVERAL QFS

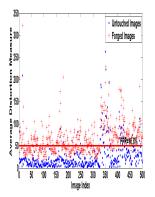
QF	50	60	70	80	90	100
Method						
MLE	38.73	37.33	37.44	37.36	37.32	34.14
Benford	59.95	58.67	58.70	58.72	58.38	80.04
Auto	9.23	11.11	11.10	11.12	11.24	8.96
Max.Peak	11.27	11.29	11.30	11.30	11.30	11.56

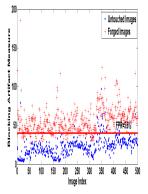
statistical modeling and nearly as fast as that autocorrelation method. However, average accuracy of our method is far higher. MLE is reliable with 83% accuracy but with more than double the time. Benford's law based method has an accuracy of 75 % but is the worst in time because recompressing the image and calculating distributions for each compressed version may become time consuming for larger images. Images used in the experiments were of size 640×480.

B. Forfery Detection

From the untouched previously compressed bitmap image set, we selected 500 images for each quality factor, each of which was subjected to four common forgeries; cropping, rotation, composition, and brightness changes. Cropping forgeries were done by deleting some columns and rows from the original image. An image was rotated by 270° for rotation forgeries Copy-paste forgeries were done by randomly copying a block of pixels from an arbitrary image and placing it in the original image. Random values were added to every pixel of the image to simulate brightness change. The resulting fake images were then stored in their uncompressed form for a total of $(500\times4)\times4=8,000$ images. Next, the quantization table for each of these images was estimated as above and used to calculate the image's average distortion, (2), and the blocking artifact, (3), measures, respectively.

Fig. 4(a) and **(b)** show values of the average distortion measure and blocking artifact measure, respectively. The scattered dots represent 500 untouched images (averaged for all quality factors for each image) while the cross marks represent 500 images from the forged dataset. As the figure





(a) Average distortion measure

(b) Blocking artifact measure

Fig. 4. Distortion measures for untouched and tampered images.

shows, values from forged images tend to cluster higher than those from untampered images. We tested the distortion measure for untouched images against several threshold values and calculated the corresponding false positive rate FPR (the number of untouched images declared as tampered), An ideal case would be a threshold giving zero false positive. However, we had to take into account the false negatives (the number of tampered images declared as untampered) that may occur when testing for forgeries. Hence, we require a threshold value keeping both FPR and the FNR low. For average distortion measure, we selected a value that gave FPR of 10.8% and a lower FNR as possible for the different types of forgeries for average distortion. The horizontal line marks this threshold $\tau =$ 50. Similarly, we selected the BAM's threshold to be $\tau = 40$, with a corresponding FPR of 5.6%. Table 4 shows the false negative rate (FNR) for the different forgeries at different quality factors for bitmaps and JPEGs. As expected, as QF increases, a better estimate of the quantization matrix of the original untampered image is obtained, and as a result the error percentage decreases. Notice how the values drop than those for JPEG file. Notice also that detection of cropping is possible when the cropping process breaks the natural JPEG grid, that it, the removed rows or columns do not fall in line with the 8×8 blocking. Similarly, when the pasted part fails to fit perfectly into the original JPEG compressed image, the distortion metric exceeds the detection threshold, and a possible composite is declared. Fig. 5 shows examples of composites. The resulting distortion measures for each composite are shown in left panel. The dark parts denote low distortion whereas brighter parts indicate high distortion values. Notice the highest values correspond to the alien part and hence mark the forged area.

TABLE IV. FORGERY DETECTION ERROR RATES FOR BITMAPS AND JPEGS

Distortion	Measure	Original	Cropping	Rotation	Compositing	Brightness
Average	JPEG	12.6%	9.2%	7.55%	8.6%	6.45%
	BMP	10.8%	3.9%	4.45%	2.0%	4.9%
BAM	JPEG	6.8%	3.3%	5.95%	3.15%	5.0%
	BMP	5.6%	1.05%	3.05%	1.25%	3.7%

IV. CONCLUSIONS

The method discussed in this paper is based on using the approximated histogram of DCT coefficients of bitmaps for extracting the image's compression history; its quantization table. Also the extracted table is used to expose image forgeries. The method proved to have practically high estimation accuracy when tested on a large set of image from different sources compared to other statistical approaches. Moreover, estimation times proved to be faster than statistical methods while maintaining very good accuracy for lower frequencies. Experimental results also showed that performance for bitmaps surpasses that of JPEGs because of their lossy nature but on the other hand, it takes more time to process a bitmap.

REFERENCES

- Hamdy S., El-Messiry H., Roushdy M. I., Kahlifa M. E., "Forgery detection in JPEG compressed images", JAR-Unpublished, 2010.
- [2] Rosenholtz R., Zakhor A., "Iterative procedures for reduction of blocking effects in transform image coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 2, pp. 91–94, Mar. 1992.
- [3] Fan Z., Eschbach R., "JEPG decompression with reduced artifacts," Proc. IS&T/SPIE Symp. Electronic Imaging: Image and Video Compression, San Jose, CA, Feb. 1994.
- [4] Fan Z., and F. Li, "Reducing artifacts in JPEG decompression by segmentation and smoothing," *Proc. IEEE Int. Conf. Image Processing*, vol. II, 1996, pp. 17–20.
- [5] Tan K. T., Ghanbari M., "Blockiness detection for MPEG-2-coded video," *IEEE Signal Process. Lett.*, vol. 7, pp. 213–215, Aug. 2000.
- [6] Minami S., Zakhor A., "An optimization approach for removing blocking effects in transform coding," *IEEE Trans. Circuits Syst. Video*

- Technol., vol. 5, pp. 74-82, Apr. 1995.
- [7] Yang Y., N Galatsanos. P., Katsaggelos A. K., "Regularized reconstruction to reduce blocking artifacts of block discrete cosine transform compressed images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 3, pp. 421–432, Dec. 1993.
- [8] Luo J., Chen C.W., Parker K. J., Huang T. S., "Artifact reduction in low bit rate dct-based image compression," *IEEE Trans. Image Process.*, vol. 5, pp. 1363–1368, 1996.
- [9] Chou J., Crouse M., Ramchandran K., "A simple algorithm for removing blocking artifacts in block-transform coded images," *IEEE Signal Process. Lett.*, vol. 5, pp. 33–35, 1998.
- [10] Fan Z., de Queiroz R. L., "Maximum likelihood estimation of jpeg quantization table in the identification of bitmap compression history", in *Proc. Int. Conf. Image Process.* '00, 10-13 Sept. 2000, 1: 948–951.
- [11] Fan Z., de Queiroz R. L., "Identification of bitmap compression history: jpeg detection and quantizer estimation", in *IEEE Trans. Image Process.*, 12(2): 230–235, February 2003.
- [12] Fu D., Shi Y.Q., Su W., "A generalized benford's law for jpeg coefficients and its applications in image forensics", in Proc. SPIE Secur., Steganography, and Watermarking of Multimed. Contents IX, vol. 6505, pp. 1L1-1L11, 2007.
- [13] Swaminathan A., Wu M., Ray Liu K. J., "Digital image forensics via intrinsic fingerprints", *IEEE Trans. Inf. Forensics Secur.*, 3(1): 101-117, March 2008.
- [14] Ye S., Sun Q., Chang E.-C., "Detection digital image forgeries by measuring inconsistencies in blocking artifacts", in *Proc. IEEE Int. Conf. Multimed. and Expo.*, July, 2007, pp. 12-15.
- [15] J. Fridrich, M. Goljan, and R. Du, "Steganalysis based on JPEG compatibility", SPIE Multimedia Systems and Applications, vol. 4518, Denver, CO, pp. 275-280, Aug. 2001.
- [16] Schaefer G., Stich M., "UCID An Uncompressed Color Image Database", School of Computing and Mathematics, Technical. Report, Nottingham Trent University, U.K., 2003.
- [17] Petkov A., Cottier S., "Image quality estimation for jpeg-compressed images without the original image", EE398 Projects - Image and Video Compression, Stanford University, March 2008.

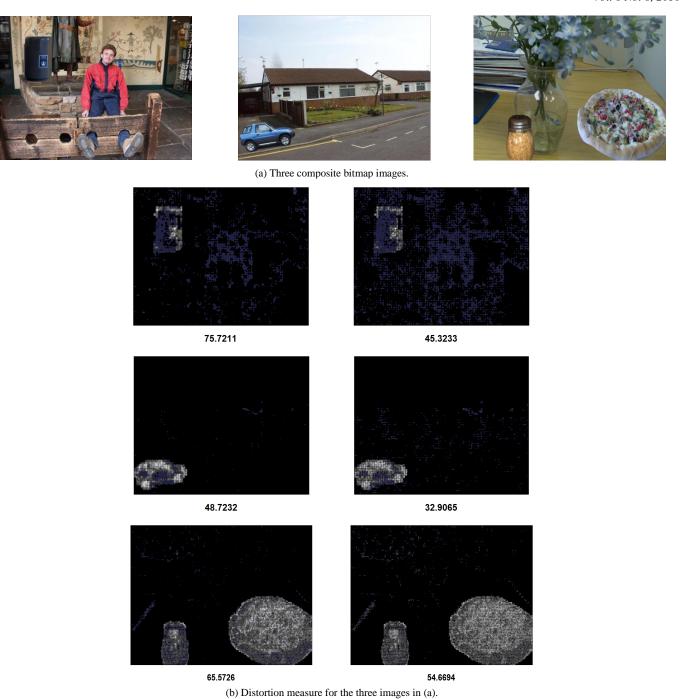


Fig. 5. Distortion measures for some composite bitmap images. The left panel represents the average distortion measure while the right panel represents the blocking artifact measure.

Steganography and Error-Correcting Codes

M.B. Ould MEDENI and El Mamoun SOUIDI Laboratory of Mathematic Informatics and Applications University Mohammed V-Agdal, Faculty of Sciences Rabat ,BP 1014, Morocco

Email: sbaimedeni@yahoo.fr, souidi@fsr.ac.ma

Abstract—In this work, We study how we used the error-correcting codes in steganographic protocols (sometimes also called the "matrix encoding"), which uses a linear code as an ingredient. Among all codes of a fixed block length and fixed dimension (and thus of a fixed information rate), an optimal code is one that makes most of the maximum length embeddable (MLE). the steganographic protocols are close in error-correcting codes. We will clarify this link, which will lead us to a bound on the maximum capacity that can have a steganographic protocol and give us a way to build up optimal steganographic protocols

Keywords: Steganography, Error-correcting code, average distortion, matrix encoding, embeding efficient.

I. Introduction

The goal of digital steganography is to modify a digital object (cover) to encode and conceal a sequence of bits (message) to facilitate covert communication. This process is fundamentally different from cryptography, where the existence of a secret message may be suspected by anyone who can observe the scrambled ciphertext while it is communicated.

A common technique in steganography is to embed the hidden message into a larger cover object (such as a digital image, for example) by slightly distorting the cover object in a way that on one hand makes it possible for the intended recipient to extract the hidden message, but on the other hand makes it very hard for everybody else to detect the distortion of the cover object (i.e., to detect the existence of the hidden message). The amount of noise that is naturally (inherently) present in the cover object determines the amount of distortion that can be introduced into the cover object before the distortion becomes detectable.

An interesting steganographic method is known as matrix encoding, introduced by Crandall [2]. Matrix encoding requires the sender and the recipient to agree in advance on a parity check matrix H, and the secret message is then extracted by the recipient as the syndrome (with respect to H) of the received cover object. This method was made popular by Westfeld [1], who incorporated a specific implementation using Hamming codes in his F5 algorithm, which can embed t bits of message in 2^t-1 cover symbols by changing, at most, one of them. Matrix encoding using linear codes (syndrome coding) is a general approach to improving embedding efficiency of steganographic schemes. The covering radius of the code corresponds to the maximal number of embedding changes needed to embed any message [8]. Steganographers,

however, are more interested in the average number of embedding changes rather than the worst case. In fact, the concept of embedding efficiency-the average number of bits embedded per embedding change-has been frequently used in steganography to compare and evaluate performance of steganographic schemes.

There are two parameters which help to evaluate the performance of a steganographic protocol $[n,k,\rho]$ over a cover message of N symbols: the average distortion $D=\frac{R_a}{N}$, where R_a is the expected number of changes over uniformly distributed messages; and the embedding rate $E=\frac{t}{N}$, which is the amount of bits that can be hidden in a cover message. In general, for the same embedding rate a method is better when the average distortion is smaller.

The remainder of the paper is organized as follows. Section 2 gives a brief introduction to Error-correcting Codes. In Section 3 we introduces the relations between error-correcting codes and steganography. We construct our approach and report on experimental results in section 4. Section 5 gives a conclusion.

II. BASICS OF CODING THEORY

We now review some elementary concepts from coding theory that are relevant for our study. A good introductory text to this subject is, for example, [7]. Throughout the text, boldface symbols denote vectors or matrices. A binary code C is any subset of the space of all n-bit column vectors $x = (x_1, ..., x_n) \in \{0, 1\}^n$. The vectors in C are called codewords. The set $\{0,1\}^n$ forms a linear vector space if we define the sum of two vectors and a multiplication of a vector by scalar using the usual arithmetics in the finite field GF(2); we will denote this field by F_2 . For any $C, D \subset F_2^n$ and vector x, $C+D=\{y\in F_2^n|y=c+d,c\in C,d\in D\}$; $x + C = \{y \in F_2^n | y = x + c, c \in C\}$. The Hamming weight w of a vector x is defined as the number of ones in x. The distance between two vectors x and y is the Hamming weight of their difference d(x,y) = w(x-y). For any $x \in C$ and a positive real number r, we denote as B(x,r) the ball with center x and radius r, $B(x,r) = \{y \in F_2^n | d(x,y) \le r\}$. We also define the distance between x and set $C \subset F_2^n$ as $d(x,C) = min_{c \in C} d(x,c)$. The covering radius R of C is defined as $R=\max_{x\in F_2^n}d(x,C)$. The average distance to code C, defined as $R_a=2^{-n}\sum_{x\in F_2^n}d(x,C)$, is the average distance between a randomly selected vector from F_2^n and the code C. Clearly $R_a \leq R$.

Linear codes are codes for which C is a linear vector subspace of F_2^n . If C has dimension k, we call C a linear code of length n and dimension k (and codimension n-k), or we say that C is an [n,k] code. Each linear code C of dimension k has a basis consisting of k vectors. Writing the basis vectors as rows of an $k \times n$ matrix G, we obtain a generator matrix of C. Each codeword can be written as a linear combination of rows from G. There are 2^k codewords in an [n,k] code. Given $x,y\in F_2^n$, we define their dot product $x.y=x_1y_1+x_2y_2+...+x_ny_n$, all operations in GF(2). We say that x and y are orthogonal if x.y=0. Given a code C, the dual code of C, denoted as C^\perp , is the set of all vectors x that are orthogonal to all vectors in C. The dual code of a [n,k] code is a [n,n-k] code with an $(n-k)\times n$ generator matrix H with the property that

$$Hx = 0 \Leftrightarrow x \in C.$$
 (1)

The matrix H is called the parity check matrix of C. For any $x \in F_2^n$, the vector s = Hx is called the syndrome of x. For each syndrome $s \in F_2^{n-k}$, the set $C(s) = \{x \in F_2^n | Hx = s\}$ is called a coset. Note that C(0) = C. Obviously, cosets associated with different syndromes are disjoint. Also, from elementary linear algebra we know that every coset can be written as C(s) = x + C, where $x \in C(s)$ arbitrary. Thus, there are 2^{n-k} disjoint cosets, each consisting of 2^k vectors. Any member of the coset C(s) with the smallest Hamming weight is called a coset leader and will be denoted as $e_L(s)$.

1) Lemma: Given a coset C(s), for any $x \in C(s)$, $d(x,C) = w(e_L(s))$. Moreover, if d(x,C) = d(x,c') for some $c' \in C$, the vector x - c' is a coset leader.

Proof: $d(x,C) = min_{c \in C}w(x-c) = min_{y \in C(s)}w(y) = w(e_L(s))$. The second equality follows from the fact that if c goes through the code C, x-c goes through all members of the coset C(s).

2) Lemma: If C is an [n,k] code with a $(n-k) \times n$ parity check matrix H and covering radius R, then any syndrome $s \in F_2^{n-k}$ can be written as a sum of at most R columns of H and R is the smallest such number. Thus, we can also define the covering radius as the maximal weight of all coset leaders.

Proof: Any $x \in F_2^n$ belongs to exactly one coset C(s) and from Lemma 1 we know that $d(x,C) = w(e_L(s))$. But the weight $w(e_L(s))$ is the smallest number of columns in H that must be added to obtain s.

III. LINEAR CODES FOR STEGANOGRAPHY

The behavior of a steganographic algorithm can be sketched in the following way:

- 1) a cover-medium is processed to extract a sequence of symbols v, sometimes called cover-data;
- 2) v is modified into s to embed the message m; s is sometimes called the stego-data;

3) modifications on *s* are translated on the cover-medium to obtain the stego-medium.

Here, we assume that the detectability of the embedding increases with the number of symbols that must be changed to go from v to s ([5] for some examples of this framework). Syndrome coding deals with this number of changes. The key idea is to use some syndrome computation to embed the message into the cover-data. In fact, such a scheme uses a linear code C, more precisely its cosets, to hide m. A word s hides the message m if s lies in a particular coset of s, related to s. Since cosets are uniquely identified by the so-called syndromes, embedding/hiding consists exactly in searching s with syndrome s, close enough to s.

A. Matrix Encoding

We first set up the notation and describe properly the matrix encoding framework and its inherent problems. Let $v \in F_q^n$ denote the cover-data and $m \in F_q^r$ the message. We are looking for two mappings, embedding Emb and extraction Ext, such that

$$\forall (v,m) \in F_q^n \times F_q^r, Ext(Emb(v,m)) = m. \tag{2}$$

$$\forall (v,m) \in F_q^n \times F_q^r, d(v, Emb(v,m)) \le T. \tag{3}$$

Equation (2) means that we want to recover the message in all cases; (3) means that we authorize the modification of at most T coordinates in the vector v.

From Error Correcting Codes (Section 2), it is quite easy to show that the scheme defined by

$$Emb(v,m) = v + D(m - E(v))$$
(4)

$$Ext(y) = E(y) = y \times H^t. \tag{5}$$

D and E mean respectively the decoding function and the function of the syndrome. enables to embed messages of length r=n-k in a cover-data of length n, while modifying at most T=R elements of the cover-data.

The parameter $\frac{n-k}{R}$ represents the embedding efficiency, that is, the number of embedded symbols per embedding changes. Linking symbols with bits is not simple, as naive solutions lead to bad results in terms of efficiency. For example, if elements of F_q are viewed as blocks of L bits, modifying a symbol roughly leads to $\frac{L}{2}$ bit flips on average and L for the worst case.

A problem raised by the matrix encoding, as presented above, is that any position in the cover-data \boldsymbol{v} can be changed. In some cases, it is more reasonable to keep some coordinates unchanged because they would produce too big artifacts in the stego-data.

1) Example: We now give an example of a protocol steganography constructed from a linear single-error-correcting code. This was also discussed, for example, in [3]. Start from the matrix

$$H = \left(\begin{array}{ccccccc} 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{array}\right)$$

whose entries are elements of F_2 . Extracting Scheme is defined

$$Ext: F_2^7 \to F_2^3$$

 $x_3 + x_5 + x_6 + x_7$. This function can be described in terms of matrix H.In fact, y_i , is the dot product of x and the i-th row of H. We claim that Ext is a extracting function of protocol steganography (7, 3, 1) Embedding Scheme for example, Ext(0,0,1,1,0,1,0) = (1,0,0). Assume y = (1,1,1). We claim that it is possible to replace x = (0, 0, 1, 1, 0, 1, 0) by x' such that Ext(x') = (1,1,1) and d(x,x') = 1. In fact, we claim more: the coordinate where x has to be changed is uniquely determined. In our case, this is coordinate number 6, so x' = (0,0,1,1,0,0,0), Here is the general embedding rule: form Ext(x) + y, (in the example this is 011). Find the column of H which has these entries (in our example, this is the sixth column). This marks the coordinate where x needs to be changed to embed payload y. This procedure indicates how H and Ext were constructed and how this can be generalized: the columns of H are simply all nonzero 3-tuples in some order. In general, we start from our choice of n and write a matrix H whose columns consist of all nonzero n-tuples. Then H has $N=2^n-1$ columns. The extracting function, $Ext: F_2^N \to F_2^n$ is defined by way of the dot products with the rows of H. Finally it is clear that embedding efficiency = 3.

IV. ASYMPTOTICLY TIGHT BOUND ON THE PERFORMANCE OF EMBEDDING SCHEMES

Our goal in this section is obtain steganographic protocols asymptotically optimal. Initially, the relationship updates in Section 3, between steganographic protocols and error-correcting codes, we used to translate the bounds on the error-correcting codes in bounds, upper and lower, of the maximum number of messages in a schema. The bounds on the error-correcting codes are known to be achieved by linear codes, we use this result to the next section to construct our protocols.

2) Proposition: [(Zhang [2], Theorem 6).]

The parameters $[n,k,\rho]$ of a steganographic protocol defined over a field F_q of cardinality q, verify that $q^k \leq V_q(n,\rho)$. **Proof.** it suffices to prove the result for proper

Proof. It suffices to prove the result for proper steganographic protocols. Take $x \in F_q^n$. For all $s \in F_q^k$ there exists $y \in B(x, \rho)$ such that Ext(y) = s, hence $card(B(x, \rho)) \ge card(F_q^k)$

This bound is analogous to the Hamming bound in Coding Theory. Thus, we can refer to it as the steganographic Hamming bound. Protocols reaching equality are called maximum length embeddable (MLE) in [2]

3) Proposition: Denote by $r_L(n,\rho)$ the largest possible value of r=n-k for $[n,k,\rho]$ steganographic protocol, similarly let $r(n,\rho)$ denote the largest possible value of r=n-log(M) or M the number of messages that can hide using binary coverwords of length n. Let us recall the following result

$$log(\sum_{i=0}^{\rho} C_n^i) - log(n) \le r_L(n,\rho) \le r(n,\rho) \le log(\sum_{i=0}^{\rho} C_n^i)$$

4) Corollary: Let h(n,T) be the maximal number of bits embeddable by schemes using binary coverwords of length n with quality threshold T. We have

$$r_L(n,T) \le h(n,T) \le r(n,T)$$

Proof. Hence, an steganographic protocol (with quality threshold T) is a slightly stronger condition than T-covering, because steganographic protocol generates not a single, but |X| disjoint T-coverings. But in particular case of linear coverings these two notions coincide.

V. CONCLUSION

Application of error-correcting codes to data embedding improves embedding efficiency and security of steganographic schemes. In this paper, we show some relations between steganographic algorithms and error-correcting codes. By using these relations we give some bound on the performance of embedding schemes

REFERENCES

- [1] R. Crandall: Some notes on steganography, *Posted on Steganography Mailing List*, **62** (1998), http://os.inf.tu-dresden.de/ westfeld/crandall.pdf.
- [2] Zhang. W, S. Li, A coding problem in steganography, Des. Codes Cryptogr, 46(2009) pp 67-81.
- [3] A. Westfeld: F5: A steganographic algorithm, High capacity despite better steganalysis. In: Moskowitz, I.S., (ed.) IH 2001 LNCS, vol. 2137, pp. 289302. Springer, Heidelberg (2001).
- [4] G. J. Simmons: The prisoners problem and the subliminal channel, in Advances in Cryptology, pp. 5167, Plenum Press, New York, NY, USA, 1984.
- [5] C. Cachin: An information-theoretic model for steganography, in D. Aucsmith (ed.): Information Hiding. 2nd International Workshop, LNCS vol. 1525, Springer-Verlag Berlin Heidelberg (1998), 306-318.
- [6] Y. Kim, Z. Duric, D. Richards: Modified matrix encoding technique for minimal distortion steganography, In: Camenisch, J.L., Collberg, C.S., Johnson, N.F., Sallee, P. (eds.) IH 2006. LNCS, vol. 4437, pp. 314-327 (2007)
- [7] F.J. Mac Williams, N. Sloane: The Theory of Error Correcting Codes, North-Holland, Amsterdam, 1977.
- [8] M.B. Medeni, EL. Souidi : A Steganography Schema and Error-Correcting Codes, Journal of Theoretical and Applied Information Technology, 7Vol18No1, pp 42-47 (2010).

A Comparative Study on Kakkot Sort and Other Sorting Methods

Rajesh Ramachandran

HOD, Department of Computer Science Naipunnya Institute of Management & Information Technology, Pongam, Kerala

Abstract: Several efficient algorithms were developed to cope with the popular task of sorting. Kakkot sort is a new variant of Quick and Insertion sort. The Kakkot sort algorithm requires $O(n \log n)$ comparisons for worst case and average case. Typically, **Kakkot Sort** is significantly faster in practice than other $O(n \log n)$ algorithms, because its inner loop can be efficiently implemented architectures. This sorting method requires data movement, but less than that of insertion sort. This data movement can be reduced by implementing the algorithm using linked list. In this comparative study the mathematical results of Kakkot sort verified experimentally on ten randomly generated unsorted numbers. To have some experimental data to sustain this comparison four different sorting methods were chosen and code was executed and execution time was noted to verify and analyze the performance. The Kakkot Sort algorithm performance was found better as compared to other sorting methods.

Key words: Complexity, performance of algorithms, sorting

Dr.E.Kirubakaran

Sr.DGM(Outsourcing),BHEL,Trichy

Introduction

Sorting is any process of arranging items in some sequence and/or in different sets, and accordingly, it has two common, yet distinct meanings:

- 1. ordering: arranging items of the same kind, class, nature, etc. in some ordered sequence,
- 2. categorizing: grouping and labeling items with similar properties together (by sorts).

In computer science and mathematics, a **Sorting Algorithm** is an algorithm that puts elements of a list in a certain order. The most-used orders are numerical order and lexicographical order. Efficient sorting is important to optimizing the use of other algorithms (such as search and merge algorithms) that require sorted lists to work correctly.

To analyze an algorithm is to determine the amount of resources (such as time and storage) necessary to execute it. Most algorithms are designed to work with inputs of arbitrary length. Usually the efficiency or complexity of an algorithm is stated as a function relating the input length to the number of steps (time complexity) or storage locations (space complexity). Algorithm analysis is an important part of a broader computational complexity theory,

which provides theoretical estimates for the resources needed by any algorithm which solves a given computational problem. These estimates provide an insight into reasonable directions of search for efficient algorithms. In theoretical analysis of algorithms it is common to estimate their complexity in the asymptotic sense, i.e., to estimate the complexity function for arbitrarily large input. Big O notation, omega notation and theta notation are used to this end

Time complexity

Time efficiency estimates depend on what we define to be a step. For the analysis to correspond usefully to the actual execution time, the time required to perform a step must be guaranteed to be bounded above by a constant. In mathematics, computer science, and related fields, Big Oh notation describes the limiting behavior of a function when the argument tends towards a particular value or infinity, usually in terms of simpler functions. Big O notation allows its users to simplify functions in order to concentrate on their growth rates: different functions with the same growth rate may be represented using the same O notation.

Although developed as a part of pure mathematics, this notation is now frequently also used in computational complexity theory to describe an algorithm's usage of computational resources: the worst case or average case running time or memory usage of an algorithm is often expressed as a function of the length of its input using big O notation.

Space complexity

The better the time complexity of an algorithm is, the faster the algorithm will carry out his work in practice. Apart from

time complexity, its space complexity is also important: This is essentially the number of memory cells which an algorithm needs. A good algorithm keeps this number as small as possible, too. The space complexity of a program (for a given input) is the number of elementary objects that this program needs to store during its execution. This number is computed with respect to the size n of the input data.

There is often a time-space-tradeoff involved in a problem, that is, it cannot be solved with few computing time and low memory consumption. One then has to make a compromise and to exchange computing time for memory consumption or vice versa, depending on which algorithm one chooses and how one parameterizes it.

In addition to varying complexity, sorting algorithms also fall into two basic categories — comparison based and non-comparison based. A comparison based algorithm orders a sorting array by weighing the value of one element against the value of other elements. Algorithms such as Quicksort, Mergesort, Heapsort, Bubble sort, and Insertion sort are comparison based. Alternatively, a noncomparison based algorithm sorts an array without consideration of pairwise data elements. Radix sort is a non-comparison based algorithm that treats the sorting elements as numbers represented in a base-M number system, and then works with individual digits of M.

Another factor which influences the performance of sorting method is the behavior pattern of the input. In computer science, best, worst and average cases of a given algorithm express what the resource usage is at least, at most and on average, respectively. Usually the resource being considered is running time, but it could also be memory or other resources.

Kakkot Sort

Kakkot Sort is a sorting algorithm that, makes O (n log n) (Big Oh notation) comparisons to sort n items. Typically, Kakkot Sort is significantly faster in practice than other O (n log n) algorithms, because its inner loop can be efficiently implemented on most architectures. This sorting method requires data movement but less than that of insertion sort. This data movement can be reduced by implementing the algorithm using linked list. Major advantage of this sorting method is its behavior pattern is same for all cases, ie time complexity of this method is same for best, average and worst case

How it sorts

From the given set of unsorted numbers, take the first two numbers and name it as key one and key two, ie, K1 and K2. Read all the remaining numbers one by one. Compare each number first with K2. If the number is greater than or equal to K2 then place the number right of K2 else compare the same number with K1. If the number is greater than K1 then place the number immediate right of K1 else left of K1.Conitnue the same process for all the remaining numbers in the list. Finally we will get three sub lists. One with numbers less than or equal to K1, one with numbers greater than or equal to K2 and the other with numbers between K1 and K2. Repeat the same process for each sub list. Continue this process till the sub list contains zero elements or one element.

Algorithm

Kakkot Sort(N:Array of Numbers, K1 ,K2 , A:integers,)

- Step1. Read the first two numbers from N, Let K1 & K2
- Step2. Sort K1 and K2
- Step3. Read the next number, Let A
- Step4. Compare A with K2
- Step5. If A is greater than or equal to K2 then place A right of K2 else compare A with K1.

 If A is less than K1 then place A left of K1 else
- Place A immediate right of K1 Step6. If the list contains any more elements go to step 3
- Step 7. Now we have 3 Sub list.
 - ✓ First list with all values less than or equal to K1.
 - ✓ Second with values between K1 and K2
 - ✓ Final with values greater than or equal to K2.

Step8. If each list contains more than 1 element go to step1

Step 9 End.

Time complexity

If there are 'n' numbers, then each iteration needs maximum 2 * (n-2) comparison and minimum of n-2 comparison and plus one. So if we take the average it will be

$$=(2n-4+n-2)/2 + 1$$

= $(3n-6)/2+1$
= $3n/2 - 2$

In the average case each list would have 3 sub lists and number of iteration will be $3^x=n$

taking logarithm on both side we get

 $x \log 3 = \log n$ $x = \log n / \log 3$ $x = \log n / 0.4771$

Ignoring the constant we can write $x = \log n$

That is there will be log n iterations and each require 3n/2 - 2 comparisons. So the time complexity of Kakkot Sort in average case is $3n/2 - 2 * \log n$. When we represent in Big Oh notation constants can be ignored, so we get $O(n \log n)$.

If the list is already in sorted order, then two comparison will be required for each number ,so total no of comparison required for each iteration will be (n-2)+1, i.e. n-1 and number of iteration will be n-1+n-3+n-6+.....+1

This can be written as 1+3+5+....n-3+n-1.

Sum of this series is S = N/2*(2a + (N-1)*d)

Where N is the number of terms in the series

'a' is first term

'd' is the difference

To get N^{th} term, the equation is a+(N-1) d

And here Nth term is n-1, so

1+(N-1)*2=n-1

2N=n

N=n/2

S=N/2(2*1+(N-1)*2)

S=N/2(2+2N-2)

S = N/2(2N)

 $Sum = N^2$

Substitute value for N we get

(n/2)**2

This is equal to one forth of n². So Kakkot Sort requires only one forth of Quick sort comparison in worst case. This is almost equal to average case time complexity. So we can say that time complexity of Kakkot sort is similar in all the cases.

Now let me manually calculate the number of comparison that Kakkot sort take.

Consider the following randomly generated ten unsorted numbers

1,60,33,3,35,21,53,19,70,94 List 1

First two numbers are 1 and 60 and sort it . Here K1 is 1 and K2 is 60

Now the total comparison is one.

Read the remaining numbers one by one

Read 33, since 33 is less than K2 and greater than K1 it need two comparison. Now the total comparison is increased to 3.

Read 3, total comparison is now 5

Read 35, total comparison is now 7

Read 21, total comparison is now 9

Read 53, total comparison now is 11

Read 19, total comparison now is 13

Read 70, total comparison now is 14

Read 94 total comparison now is 15

Now the list will be

1, 3,35,21,53,19,**60**,70,94

Here we have 3 sublist

The first one with zero elements

Second list is , 3,35, 35,21,53,19

Third list is 70,94

Now do the same process second and third

Second list

Read first two numbers, and sort

We have K1 = 3 and K2 = 35

Now total comparison is 16

Read 21, total comparison now is 18

Read 53,total comparison now is 19

Read 19,total comparison now is 21

Now the list will be

3.19.21.**35**.53

Now only one list with more than one element, ie 19 and 21

Read the first two numbers and sort

Here K1=19 and K2=21

Now the total comparison is 22

Now regarding the sublist 3 we have two numbers 70 and 94

Read the numbers and sort

Now the total number of comparison is 23

So Using Kakkot sort, to sort the given ten randomly generated numbers require only 23 comparisons.

Kakkot Sort and Qucick Sort

Time complexity of Quick sort is O(n log n) in the case of average case and O(n²) in the worst case behavior. From this it is clear that Kakkot sort is better than quick sort. While sorting Quick sort does not require any data movement where as Kakkot sort needs data movement when the item is less than first key element and greater than second key element. But this data movement can be avoided by implementing the algorithm using linked list.

To sort the above ten numbers in the List 1, Quick sort requires 29 comparisons

Kakkot Sort and Heap Sort

Heapsort is a much more efficient version of selection sort. It also works by determining the largest (or smallest) element of the list, placing that at the end (or beginning) of the list, then continuing with the rest of the list, but accomplishes this task efficiently by using a data structure called a heap, a special type of binary tree. Once the data list has been made into a heap, the root node is guaranteed to be the largest(or smallest) element. When it is removed and placed at the end of the list, the heap is rearranged so the largest element remaining moves to the root. Using the heap, finding the next largest element takes O(log n) time, instead of O(n) for a linear scan as in simple selection sort. This allows Heapsort to run in O(n log n) time, and this is also the worst case complexity.

With the same set of unsorted numbers in the List 1, Heap sort requires 30 comparisons

Kakkot Sort and Bubble Sort

Bubble sort is a straightforward and simplistic method of sorting data that is used in computer science education. algorithm starts at the beginning of the data set. It compares the first two elements, and if the first is greater than the second, then it swaps them. It continues doing this for each pair of adjacent elements to the end of the data set. It then starts again with the first two elements, repeating until no swaps have occurred on the last pass. This algorithm is highly inefficient, and is rarely used[citation needed][dubious - discuss], except as a simplistic example. For example, if we have 100 elements then the total number of comparisons will be 10000. A slightly better variant, cocktail sort, works by inverting the ordering criteria and the pass direction on alternating passes. The modified Bubble sort will stop 1 shorter each time through the loop, so the total number of comparisons for 100 elements will be 4950.

Bubble sort average case and worst case are both $O(n^2)$

For the above unsorted numbers in the List 1 Bubble sort requires 45 comparisons.

Kakkot Sort and Insertion Sort

Insertion sort is a simple sorting algorithm that is relatively efficient for small lists and mostly-sorted lists, and often is used as part of more sophisticated algorithms. It works by taking elements from the list one by one and inserting them in their correct position into a new sorted list. In arrays, the new list and the remaining elements can share the array's space, but insertion is expensive, requiring shifting all following elements over by one. Shell sort (see below) is a

variant of insertion sort that is more efficient for larger lists.

Insertion sort requires 38 comparisons to sort the above ten randomly generated numbers in the List 1.

Conclusion

From the above examples it is clear that Kakkot Sort time complexity is better than other sorting methods. Even though Kakkot sort requires data movement of items when the item is less than the key K2 and greater than the key K1, this data movement can be reduced by implementing the algorithm using linked list.

References:

- [1] Aaron M Tanenbaum, Moshe J Augenstein, "Data Structures using C", Prentice Hall International Inc., Emglewood Cliffs, NJ, 1986
- [2] Robert L Cruse, " *Data Structure and Program Design*", Prentice Hall India 3rd ed.,1999
- [3] Robert Kruse, C L Tondo, Bruse Leung "Data Structures and Program design in C", Pearson Education,2nd Ed.,2002
- [4] Alfred V Aho, John E Hopcroft, Jeffrey D Ullman, " *The Design and Analysis of Computer Alogorithms*", Pearson Education , 2003
- [5] Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, Clifford Stein, "Introduction to Algorithms" Prentice Hall of India Pvt.Ltd., 2nd Ed., 2004

- [6] Sartaj Sahni, "Data Structures Algorithms and Applications in C++", University Press, 2nd Ed., 2005
- [7] Yedidyah Langsam, Moshe J Augenstein, Aaron M Tanenbaum "*Data Structures using C and C++*", Prentice Hall India, 2nd Ed. 2005
- [8] Alfred V Aho, John E Hopcroft, Jeffrey D Ullman," Data Structures and Algorithms", Pearson Education, 2nd Ed., 2006
- [9] Sara Baase, Allen Van Gelder, "Computer Algorithms Introduction to Design and Analysis, Pearson Education, 3rd Ed. ,2006
- [10] Mark Allen Weiss "Data Structures and Algorithm analysis in C++ ", Pearson Education, 3rd Ed., 2007
- [11] Michael T Goodrich, Roberto Tamassia, "Algorithm Design Foundations, Analysis and Internet Examples", John Wiley and Sons Inc.,2007
- [12] Seymour Lipschutz, GAV Pai, " *Data Structures*", Tata McGraw Hill,2007
- [13] Robert Lafore," Data Structures and Algorithms in Java", Waite Group Inc., 2007
- [14] Rajesh Ramachandran, Dr.E. Kirubakaran, "*Kakkot Sort A New Sorting Method*", International Journal of Computer Science, Systems Engineering and Information Technology, ISSN 0974-5807 Vol. 2 No. 2 pp209-213,2010

A Generalization of the PVD Steganographic Method

M.B. Ould MEDENI and El Mamoun SOUIDI

Laboratory of Mathematic Informatics and Applications University Mohammed V-Agdal, Faculty of Sciences Rabat ,BP 1014, Morocco

Email: sbaimedeni@yahoo.fr, souidi@fsr.ac.ma

Abstract—In this work we propose a novel Steganographic method for hiding information within the spatial domain of the gray scale image. The proposed approach works by dividing the cover into blocks of equal sizes and then embeds the message in the edge of the block depending on the number of ones in left four bits of the pixel. The purpose of this work is to generalize the PVD method [7] With four-pixel differencing instead of two-pixel differencing and use the LSB Substitution to hide the secret message in the cover image

Keywords: Steganography, Watermarking, Least Significant Bit(LSB), PVD method, Digital Images, Information Hiding, Pixel-value differencing.

I. Introduction

Steganography is the art of stealth communication. Its purpose is to make communication undetectable. The steganography problem is also known as the prisoners' dilemma formulated by Simmons [4]. Alice and Bob are imprisoned and want to hatch an escape plan. They are allowed to communicate via a channel monitored by a warden. If the warden finds out that they are communicating secretly, he throws them into solitary confinement. Thus, the prisoners need to design a method to exchange messages without raising the warden's suspicion. The prisoners hide their messages in innocuous-looking cover objects by slightly modifying them (obtaining stego objects). The embedding process is usually driven by a stego key, which is a secret shared between Alice and Bob. It is typically used to select a subset of the cover object and the order in which the cover object elements are visited during embedding. The most important property of any steganographic communication is statistical undetectability. In other words, the warden should not be able to distinguish between cover and stego objects. Formal description of this requirement in information-theoretic terms was given by Cachin [5]. If the communication channel that Alice and Bob use is distortion-free, we speak about the passive warden scenario.

The most common and well-known steganographic method is called least significant bit (LSB) substitution, which embeds secret data by replacing k LSBs of a pixel with k secret bits directly [1]. Many optimized LSB methods have been proposed to improve this work [2], [3]. The human perceptibility has a property that it is sensitive to some changes in the pixels of the smooth areas, while it is not sensitive to changes in the edge areas. Not all pixels in a cover image can tolerate

equal amount of changes without causing noticeable distortion. Hence, to improve the quality of stego images, several adaptive methods have been proposed in which the amount of bits to be embedded in each pixel is variable. Wu and Tsai proposed a novel steganographic method that uses the difference value between two neighboring pixels to determine how many secret bits should be embedded [7].

In contrary: Steganalysis methods attempt to detect Stegoimage and extract it. Inserting secret bits in image changes some statistics of image, this opens some roads to detect Stegoimage. So the changes made by Steganographic are a key performance metric; lower change: more robust algorithm. It is evident that the changes in cover image are related to the volume of inserted bit, so Stego-images with higher insertion rate are detected more easily.

Stegananalysis methods generally are divided in two main groups: active and passive methods. In passive methods only presence or absence of hidden data is considered, while in active methods a inserted data is extracted [8]. Furthermore, different steganalysis methods, depending on steganography algorithms they target, can be classified in two groups: Modelbased (Specific) and Universal Steganalysis.

The aim of this work is to generalize the PVD method [7] With four-pixel differencing instead of two-pixel differencing and LSB Substitution. The remainder of the paper is organized as follows. Section 2 gives a brief introduction to Steganography and Data Hiding Methods. We construct our approach and report on experimental results in section 3 and 4. Section 5 gives a conclusion.

II. DIGITAL IMAGES IN STEGANOGRAPHY

A. Digital Images

A digital image at the most abstract level is a two-dimensional array of colored pixels or dots. When these pixels are displayed on a high-resolution monitor and viewed at an appropriate distance, they appear to be a continuously colored image. Each pixel is a certain color which is typically defined, using the redgreen- blue (RGB) color model, as a combination of varying amounts of red, green, and blue light. A color image is therefore said to contain three bands, each of which represents the amount of red, green, or blue light in the image. Whereas a color image contains color and intensity information, a gray-scale image is composed of

pixels that vary only in intensity, not color. Gray-scale images therefore have only a single band. Without loss of generality, the remaining discussion will focus on gray-scale images. The discussion is easily extended to cover color images by noting that a color image is the composition of three individual gray-scale images representing the red, green and blue bands. The typical gray-scale image has an 8-bit depth which is sufficient to represent 256 unique intensity values ranging from black to white [9]. A brief review of binary representation will be instructive when interpreting bit-level pixel data in the context of a digital image. An 8-bit binary numeral has the general form

$$A_72^7 + A_62^6 + \dots + A_12^1 + A_02^0$$

where A_n represents a single binary digit. In a digital image it is clear that A_7 is the most significant bit and indicates whether the pixel value is greater than 127. A common means of converting a grayscale image to a binary (i.e. black-and-white) image is to extract the A_7 bit from each pixel. By contrast, A_0 embodies relatively little information and, in the context of a digital image, can generally be understood as a noise channel.

B. Overview of Steganograhy

Steganography hides secret messages under the cover of a carrier signal so it cannot be seen or detected [6], [8], [11]. Steganography technique should generally possess two important properties: good visual/statistical imperceptibility and a sufficient payload. The first is essential for the security of hidden communication and the second ensures that a large quantity of data can be conveyed [10]. Two levels of protection can be done if the message is encrypted before hiding it, so it must be decrypted before reading it. Invisible watermarking is treated as a subset of steganography [10].

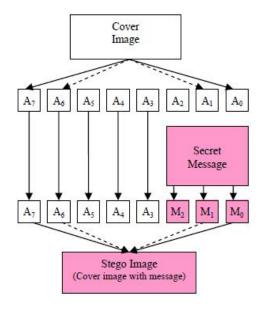


Fig. 1. Embedding of a secret message into the three least significant channels of a cover image

The difference is that steganography conceals a message so that this hidden message is the object of the communication where in watermarking; the hidden message provides important information about the cover media, such as authentication or copyright. Steganography, in the simplest case, capitalizes on this overabundance of information by replacing the noise channels (i.e. the least significant bit channels) with an arbitrary secret message. Figure 1 gives an overview of a steganographic process flow. A source image, hereafter referred to as a cover, is viewed as 8 information carrying channels. A secret message is spread over the least significant channels (in this case the three least significant channels) with the modified channels re-combined to obtain an output, hereafter referred to as the stego image, that visually resembles the cover image and contains the injected message.

III. PVD METHOD FOR GRAY-LEVEL IMAGE

The pixel-value differencing (PVD) method [7] segments the cover image into nonoverlapping blocks containing two connecting pixels and modifies the pixel difference in each block (pair) for data embedding. A larger difference in the original pixel values allows a greater modification. The hiding algorithm is described as follows:

- 1) Calculate the difference value d_i for each block of two consecutive pixels P_i and P_{i+1} , $d_i = P_{i+1} P_i$
- 2) Find the optimal R_i of the d_i such that $R_i = min(u_i k)$, where $u_i \ge k$, $k = |d_i|$ and $R_i \in [l_i, u_i]$
- 3) Decide t bits of secret data which are hidden with each d_i , i.e. each block of two consecutive pixels is defined as $t = log_2(w_i)$ where w_i is the width of the R_i
- 4) Read t bits binary secret data one by one according to Step 3, and then transform t into decimal value b. For instance, assume a binary secret data is 101, then b=5.
- 5) Calculate the new difference value d_i' using: $d_i' = l_i + b$, for $d_i \ge 0$ or $d_i' = -(l_i + b)$, for $d_i < 0$
- 6) P_i and P_{i+1} are modified to hide t secret data by the following formula: $(P_i', P_{i+1}') = (P_i \lceil \frac{m}{2} \rceil, P_{i+1} + \lfloor \frac{m}{2} \rfloor)$: $d_i \in odd$ or $(P_i', P_{i+1}') = (P_i \lfloor \frac{m}{2} \rfloor, P_{i+1} + \lceil \frac{m}{2} \rceil)$: $d_i \in even$ where $m = d_i' d_i$. Finally, we compute the values of (P_i', P_{i+1}') which represent the secret data.
- 7) Repeat Steps 1-6, until all secret data are hidden into the cover image and the stego-image is obtained.

In the extraction phase, the original range table is necessary. It is used to partition the stego-image by the same method as used to the cover image. The extraction phase is implemented as follows:

- Calculate the difference value d'_i between each two successive pixels for each block (P'_i, P'_{i+1}) from the following formula : d'_i = |P'_{i+1} P'_i|
 Find the optimum R_i of the d'_i just as in Step 2 in the
- 2) Find the optimum R_i of the d'_i just as in Step 2 in the hiding phase.
- 3) Obtain b' by subtracting l_i from d'_i . The b' value represents the value of the secret data in decimal.
- 4) Convert b' into binary then find number of bits t from the secret data, where $t = log_2(w_i)$ [7]

IV. PROPOSED STEGANOGRAPHY SCHEME

In this section we discuss the proposed approach for hiding information within the spatial domain of the gray scale image. The proposed approach works by dividing the cover into blocks of equal sizes (8×8) . Our proposed method adaptively embeds messages using two levels (lower-level and higherlevel), and the square of median value M is used to partition the average difference D into two levels. If D < M, D belongs to "lower-level" (i.e., the block belongs to a smooth area). Otherwise, D belongs to "higher-level" (i.e., the block belongs to an edge area).

A. Determine The Place of Embedding in The Image

All the pixels in the cover image are 256 gray values. The cover image is partitioned into non-overlapping four-pixel blocks. For each block, there are four neighboring pixels $p_{i,j}$, $p_{i,j+1}, p_{i+1,j}, p_{i+1,j+1}$, and their corresponding gray values are y_1 , y_2 , y_3 and y_4 , respectively.

- 1) Divide the cover into blocks of equal sizes 8×8
- 2) Calculate the square root of median for each block. M = $\sqrt{(median)}$
- 3) Calculate the average difference value D, which is given
- by $D = \frac{1}{3} \sum_{i=0}^{3} (y_{i+1} y_i)$ 4) IF $D \ge M$, then embed Message in $p_{i,j}, p_{i,j+1}, p_{i+1,j}$, $p_{i+1,j+1}$, (go to The embedding algorithm)

B. The embedding algorithm

- 1) Split each pixel into two equal parts (see Figure 2).
- 2) Count number of 1 in the most part and embed a secret message in the least part according to the corresponding number of bits in Table 1.

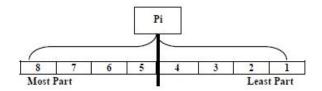


Fig. 2. Split Process.

number of 1 in	number of Bits
the most part	to be embedded
4 or 3	3 bits
2	2 bits
1 or 0	1

The recipient uses the extraction algorithm in order to extract the secret message from the stego-image. Extracting secret message is done in the same way as in the embedded operation, depending on the value of the median: M = $\sqrt{(median)}$. If the average difference value D is more than the value of M then extract the message depending on the rule in Table 1.

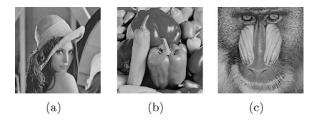
V. EXPERIMENTAL RESULTS

Several experiments are preformed to evaluate our proposed method. Ten gray-scale images with size 512×512 are used in the experiments as cover images, and three of them are shown in Fig. 3. A series of pseudo-random numbers as the secret bit streams are embedded into the cover images. The peak signal to noise ratio (PSNR) is utilized to evaluate the quality of the stego image. For an $M \times N$ gray-scale image, the PSNR value is defined as follows:

$$PSNR = 10 \times log_{10} \frac{255 \times 255 \times M \times N}{\sum_{i=1}^{M} \sum_{j=1}^{N} (P_{ij} - Q_{ij})^2} (dB)$$

where P_{ij} and Q_{ij} denote the pixel values in row i and column j of the cover image and the stego image, respectively. In this section we present the experimental results of stego-image on three will known images: Lena, Pepper, and Baboon images. These images are shown in Figs 3. The quality of stego-image created by our proposed method are shown in Figs.4. As the figures show, distortions resulted from embedding are imperceptible to human vision. We present also a comparative study of the proposed methods with PVD method.

We have analyzed our results according to PVD method for each of the tested images. We also analyzed our results by computing Payload, and peak signal-to noise ratio (PSNR).



Three cover images with size 512×512 : (a) Lena (b) Peppers (c) Fig. 3.

Payload: the size of date that could be imbedded within the cover-image is shown in Table 2

Image	Image size	Data size	Data size
		(PVD)	(Proposed Method)
Lena	128×128	2048	2493
	255×255	8192	10007
	512×512	32768	40017
	1024×1024	131072	160604
Peppers	128×128	2048	2560
	255×255	8192	10211
	512×512	32768	40990
	1024×1024	131072	163724
Baboon	128×128	2048	2443
	255×255	8192	9767
	512×512	32768	39034
	1024×1024	131072	156308

Figure. 4 shows the amount of messages hidden in the 3 cover images. Three stego images (a) Lena (embedded 40017 bits, PSNR = 42.68dB) (b) Peppers (embedded 40990

bits, PSNR = 43.23dB) (c) Baboon (embedded 39034 bits, PSNR = 37.71dB).







Fig. 4. Three stego images: (a) Lena (b) Peppers (c) Baboon.

VI. CONCLUSION

In this paper, we have proposed a novel steganographic method based on four-pixel differencing and LSB substitution. Secret data are hidden into each pixel by the k-bit LSB substitution method, where k is decided by the number of 1 in the most part for pixel. Experimental results showed that the proposed method gave best values for the PSNR measure, which means that there is no difference between the original, and the stegano-images.

REFERENCES

- [1] D.W. Bender, N.M. Gruhl, A. Lu, : Techniques for data hiding, *IBM Syst. J.* 35 (1996) 313-316
- [2] R.Z. Wang, C.F. Lin, J.C. Lin, *Image hiding by optimal LSB substitution and genetic algorithm*, Pattern Recognit. 34 (3) (2001) 671-683.
- [3] I.C. Lin, Y.B. Lin, C.M. Wang,: , Hiding data in spatial domain images with distortion tolerance, Comput. Stand. Inter. 31 (2) (2009) 458-464.
- [4] G. J. Simmons: The prisoners problem and the subliminal channel, in Advances in Cryptology, pp. 5167, Plenum Press, New York, NY, USA, 1984
- [5] C. Cachin: An information-theoretic model for steganography, in D. Aucsmith (ed.): Information Hiding. 2nd International Workshop, LNCS vol. 1525, Springer-Verlag Berlin Heidelberg (1998), 306-318.
- [6] Y. Kim, Z. Duric, D. Richards: Modified matrix encoding technique for minimal distortion steganography, In: Camenisch, J.L., Collberg, C.S., Johnson, N.F., Sallee, P. (eds.) IH 2006. LNCS, vol. 4437, pp. 314-327 (2007).
- [7] D. C. Wu and W. H. Tsai, : A steganographic method for images by pixelvalue differencing, Pattern Recognition Letters, 24(9-10), pp.16131626, 2003.
- [8] A. Westfeld and A. Pfitzmann: Attacks on Steganographic Systems, 3rd International Workshop. Lecture Notes in Computer Science, Vol.1768. Springer-Verlag, Berlin Heidelberg New York (2000) 61-75.
- [9] Kenny Hunt: A Java Framework for Experimentation with Steganography, ACM SIGCSE Bulletin Proceedings Volume 37 Issue 1, 2005. pp.282-286
- [10] C. C. Chang, W. L. Tai, and C. C. Lin: A novel image hiding scheme based on VQ and Hamming distance, Fundamenta Informaticae, vol. 77, no. 3, pp. 217-228, 2007.
- [11] M. B. Medeni and El. Souidi: A Novel Steganographic Protocol from Error-correcting Codes, Journal of Information Hiding and Multimedia Signal Processing, Volume 1, Number 4, October 2010, pp 337-343...

Implementation of Polynomial Neural Network in Web Usage Mining

S.Santhi

Research Scholar Mother Teresa Women's University Kodaikanal, India Dr. S. Purushothaman
Principal
Sun college of Engineering and Technology
Nagarkoil, India

Abstract-Education, banking, various business and humans' necessary needs are made available on the Internet. Day by day number of users and service providers of these facilities are exponentially growing up. The people face the challenges of how to reach their target among the enormous Information on web on the other side the owners of web site striving to retain their visitors among their competitors. Personalized attention on a user is one of the best solutions to meet the challenges. Thousands of papers have been published about Most of the papers are distinct either in personalization. gathering users' logs, or preprocessing the web logs or Mining In this paper simple codification is performed to filter the valid web logs. The codified logs are preprocessed with polynomial vector preprocessing and then trained with Back Propagation Algorithms. The computational efforts are calculated with various set of usage logs. The results are proved the goodness of the algorithm than the conventional methods.

Keywords- web usage mining; Back propagation algorithm;, Polynomial vector processing

I. INTRODUCTION

Web users feel comfortable if they reached the desired web page within the minimum navigation on a web site. A study of Users' recent behavior on the web will be useful to predict their desired target page. Generally Users' browsing patterns are stored in the web logs of a web server. These patterns are learned through the efficient algorithms to find the target page. Backpropagation Algorithm with Polynomial Vector Preprocessing,(BPAPVP) is implemented for learning the patterns. With learned knowledge, various set of users' browsing patterns are tested. The results are observed and presented as an analysis on computational efforts of the algorithm. The analysis on the results proves the correctness of the algorithm. Thus the BPAPVP leads to improved web usage mining than the numerous conventional methods.

A. Literature Review

Michael Chau et al. [1] attempted to use Hopfield Net for web analysis. The web structure and content analysis are incorporate into the network through a new design of network Their algorithm performed (70% of accuracy) better than traditional web search algorithms such as

breadth-first search(42.6% of accuracy) and best-first search algorithms(48.2% of accuracy). David Martens et al. [2] proposed a new active learning based approach (ALBA) to extract comprehensible rules from opaque SVM models. They applied ALBA on several publicly available data sets and confirmed its predictive accuracy. Dilhan Perera [3] et al. have performed mining of data collected from students working in teams and using an online collaboration tool in a one-semester software development project. Clustering was applied to find both groups of similar teams and similar individual members, and sequential pattern mining was used to extract sequences of frequent events. The results revealed interesting patterns characterizing the work of stronger and weaker students. Key results point to the value of analysis based on each resource and on individuals, rather than just the group level. They also found that some key measures can be mined from early data, in time for these to be used by facilitators as well as individuals in the groups. Some of the patterns are specific for their context (i.e., the course requirements and tool used). Others are more generic and consistent with psychological theories of group work, e.g., the importance of group interaction and leadership for success. Edmond H.Wu et al.[4] introduced an integrated data warehousing and data mining framework for website management. The model focuses on the page, user and time attributes to form a multidimensional can be which can be frequently updated and queried. The experiment shown that data model is effective and flexible for different analysis tasks. Gaungbin Huang et al. [5] proposed a simple learning algorithm capable of real-time learning, which can automatically determine the parameters of the network at one time only. This learning algorithm is compared with BP and k-NN algorithm. There are 4601 instances and each instance has 57 attributes. In the simulation 3000 randomly selected instances compose the training set and all the rest are used for testing. RLA achieves good testing accuracy at very fast learning speed; however BP need to spend 4641.9s on learning which is not realistic in such a practical real-time application. In the forest typed prediction problem 100,000 training data and 481012

testing data have been taken. The testing time of k-NN can be as long as 26 hours, where as RLA finished within 65.648 seconds. Incorporating neural network (NN) into supervised learning classifier system (UCS) [6] offers a good compromise between compactness, expressiveness, and accuracy. A simple artificial NN is used as the classifier's action and obtained a more compact population size, better generalization and the same or better accuracy while maintaining a reasonable level of expressiveness negative correlation learning (NCL) is also applied during the training of the resultant NN ensemble. NCL is shown to improve the generalization of the ensemble. Hongiun Lu et al.[7] proposed an neural network to extract concise symbolic rules with high They have been improving the speed of accuracy. network training by developing fast algorithms, the time required to extract rules by our neural network approach is still longer than the time needed by the decision tree approach. They tried to reduce the training time and improve the classification accuracy is to reduce the number of input units by feature selection. caverlee et al.[8] presented the Thor framework for sampling, locating and partitioning the QA-Pagelets (Query-Answer pagelets) from the Deep web. [Large and growing collection of web accessible databases known as the deep web] Their experiments have shown that the proposed page clustering algorithm achieves low-entropy clusters and the sub-tree clustering algorithms identify QA-Pagelets with excellent precision and recall. Lotfi Ben Romdhane [9] extends a neural model for casual reasoning to mechanize the monotonic class. developed Unified Neural Explainer (UNEX) for casual reasoning (independent, incompatibility and open). UNEX is mechanized by the use of Fuzzy AND-ing networks, whose activation is based on new principle. called softmin. They considered a battery of 1000 random manifestations/cases. UNEX had a coverage ration greater than 0.95 in 220 cases (22%). Magdalini Eirinaki et al. [10] presented a survey of the use of web mining for web personalization. A review of the most common methods that are used as well as technical issues that occur is given, along with a brief overview of the most popular tools and applications available from S/W vendors. Mankuan Vai et al.[11] developed a systematic approach that creates a Hop field network to represent qualitative knowledge about a system for analysis and reasoning. A simple sic node neural network is designed as a building block to capture basic qualitative relations. The objective of the transistor modelling technique is to determine the topology of an equivalent circuit and to extract its element values from the measured device data. The ultimate advantage of the neural network is in its capability of implementing the neural network as a parallel distributed processor, which will remove the time consuming factor of sequentially updating individual neurons. C. Porcel et al. [12] presented a new fuzzy

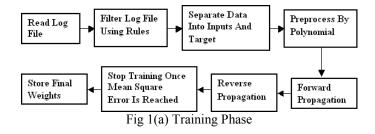
linguistic recommender system that facilitates the acquisition of the user preferences to characterize the user profiles. They allowed users to provide their preferences by means of incomplete fuzzy linguistic preference relation. The user profile is completed with user preferences on the collaboration possibilities with other users. Therefore, this recommender system acts as a decision support system that makes decisions about both the resources that could be interesting for a researcher and his/her collaboration possibilities with other researchers to form interesting working groups. The experimental results the user satisfaction with the received shown recommendations. The average of precision, recall and F1 (F1 is a combination metric that gives equal weight to both precision) metrics are 67.50%, 61.39% and 63.51%, Ranieri Barglia et al.[13] proposed a recommender system that helps user to navigate through the web by providing dynamically generated links to pages that have not been visited and are of potential interest. They contributed and suggest, a privacy enhanced recommender system that allows for creating serendipity recommendations without breaching users privacy. They said that a system is privacy safe if the two conditions hold: (i) The user activity cannot be tracked (ii) The user activity cannot be inferred. They conducted a set of experiments assess the quality of recommendations Sankar K.Pal et al. [14] summarized the different type of web mining and its basic components, along with their current states of are. The limitations of existing web mining methods / tools are explained. The relevance of soft computing is illustrated through example and diagrams. Tianyi et al. [15] is examined the problem of optimal partitioning of customer bases into homogeneous segments for building better customer profiles and have presented the direct grouping approach as a solution. That approach partitions the customers not based on computed statistics and particular clustering algorithms, but in terms of directly combining transactional data of several customers and building a single model of customer behaviour on that combined data. They formulated the optimal partitioning problem as a combinatorial optimization problem and showed that it is NP-hard. Then, three suboptimal polynomial-time direct grouping methods, Iterative Merge (IM), Iterative Growth (IG), and Iterative Reduction (IR) are shown that the IM method provides the best performance among them. It is shown that the best direct grouping method significantly dominates the statistics-based and one-to-one approaches across most of the experimental conditions, while still being computationally tractable. It is also shown that the distribution of the sizes of customer segments generated by the best direct grouping method follows a power law distribution and that micro segmentation provides the best approach to personalization. Vir V.Phoha et al.[16] developed a new learning algorithm for fast web page allocation on a server using the self-organizing properties

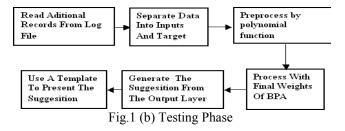
of the neural network (NN). They compared the performance of the algorithm with round-robin (RR). As the number of input objects increases, the algorithm achieves a hit ratio close to 0.98 whereas RR schema never achieve more than 0.4. Xiaozhe Wang et al.[17] proposed a concurrent neuro-fuzzy model to discover and analyze useful knowledge from the available web log data. They made use of the cluster information generated by self organizing map for pattern analysis and a fuzzy inference system to capture the chaotic trend to provide short-term(hourly) and long-term (daily) web traffic trend predictions. Yu-Hui et al.[18] explored a new data source called intentional browsing data (IBD)for potentially improving the effectiveness of WUM applications IBD is a category of online browsing actions such as "copy", "scroll", or "save as " and is not recorded in web log files. Consequently this research aims to build a basic understanding of IBD, which will lead to its easy adoption in WUM research and practice. Specially, this paper formally defines IBD and clarifies its relationship with other browsing data. Zhicheng Douet al. [19] developed an evaluation framework based on real query logs to enable large-scale evaluation of personalized search. They have taken 5 algorithms for evaluation research (i) Click-based algorithm (P-Click), (ii) longterm user topic interests (L-Topic) (iii) Short-term interests (S-Topic) (iv) Hybrid of L-Topic and S-Topic, (LS-Topic).(v) Group base personalization (G-Click). They found that no personalization algorithms can outperform others for all queries and concluded that different methods have different strength and weakness. Zi Lu et al. [20] reviewed related research results in this area and their practical significance for a comprehensive explanation of various effect functions based on utility theory. They used the data on Internet development in China and related intelligent decision models to calculate the effect function. Based on the findings, they explained the features of the effect of website information flow on realistic human flow from various aspects. Research results showed that the effect of website information flow can be divided into substitution and enhancement, so that the relationship of the website information flow in guiding the human flow changes from one dimension to multi-dimensional morphology. They indicated that, on one hand, website information flow is lagged to some extent, but is enhanced gradually and grows faster than realistic human flow; on the other hand, by comparing the evolution trend of the intensity of the two functions, it can be seen that the enhancement function occurs later than the substitution, but develops faster and has greater force. Following comparison between the simulation value and the actual value, it is proved that the effect of website information flow is basically in line with the relationship of realistic human flow. These results can support government and business in making decisions on web information publication. Through the comparison between enhancement effect and substitution effect, they found that the substitution and enhancement effect of website information flow to realistic human flow exist simultaneously. The development trend of the enhancement effect is quicker than that of the substitution effect, and the enhancement effect is stronger. The information flow guiding human flow in the initial period of the network economy suggests that the substitution effect is stronger, and in the later period that the enhancement effect is stronger and quicker.

II. PROBLEM DEFINITION

Users' browsing patterns are gathered from the web server and then extracts only the valid logs i,e., The logs that doesn't contain robots.txt, .jpg, ,gif etc and unsuccessful request. These logs are codified with Meta data of the web site. Then the codified patterns are applied to the polynomial vector for preprocessing. The preprocessed data are fed to back propagation algorithm for training the usage patterns.

Machine learning theory based web usage mining assumes no statistical information about the web logs. This work falls under the category of supervised learning by employing two phase strategies such as a) Training phase b) Testing phase. In training phase, original logs are codified by simple substitution of unique page id instead of page name for all the successful html requests and are interpolate by preprocessing into polynomial vector. The n dimensional patterns are innerproduct to obtain 2 dimensional vectors which is trained by neural classifier to learn the nature of the logs. BPA takes the role of neural classifier in this work. By training the classifier for a specific users' logs a reasonably accurate suggestions can be derive. In testing phase, various users' logs are supplied to the trained classifier to decide which page-id is to be suggested. The flow charts of both phases are given in Figure 1.a and Figure 1.b





III. IMPLEMENTATION

The simulation of personalization through web usage mining has been implemented using MATLAB 7®. Sample sets of logs are taken from ProtechSC's web server. These logs are filtered and codified. Table II gives sample codified logs that have been obtained after codification of the extended log format. Each number refers to a webpage. The % symbol is the comment and the number after the percent is the line number. Users' 50 days patterns have been collected. 25 patterns have been used for training and the remaining patterns used for testing.

A. Filter the Log File

the web logs are collected from the web server of www.protechsc.net. Sample web log file of this site is given in Fig.2

TABLE I- CODIFICATION TABLE OF WWW.PROTECHSC.NET

PageName	Code
index.html	1
aboutus.html	2
Dissertation.html	3
Whatwedo.html	4
Projecttopics.html	5
Services.html	6
consultation.html	7
Contactus.html	8
PaymentDetails.html	9
Enquies&Comment.html	10
Algorithm	11
Flowchart	12
Submit	13
SpeechSeparation.html	14
WaveletPackett.html	15
PwdAuthentication.html	16
OFDM_Frequency.html	17
CharRecog.html	18
CarotidArtery.html	19
AnalysisMRI.html	20
BPA Char.html	21
DirectSearch.html	22
Detect_micro classfication.html	23
Cloud_Contamination.html	24
Info_retrieval	25

26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52

66.249.71.171 - - [25/May/2009:18:50:11 +0530] "GET/robots.txt HTTP/1.1" 404 - "-" "Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)"

66.249.71.171 - - [25/May/2009:18:50:12 +0530] "GET/consultation.html HTTP/1.1" 304 - "-" "Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)"

66.249.71.171 - - [25/May/2009:19:05:19 +0530] "GET/dissertation.html HTTP/1.1" 304 - "-" "Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)"

66.249.71.171 -- [25/May/2009:21:43:07 +0530] "GET/Contact_us.php HTTP/1.1" 200 41264 "-" "Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)"

72.20.109.34 - - [25/May/2009:23:01:41 +0530] "GET /robots.txt HTTP/1.1" 404 - "-"
"Mozilla/5.0 (compatible; GurujiBot/1.0; +http://www.guruji.com/en/WebmasterFAQ.html)"

72.20.109.34 - - [25/May/2009:23:01:41 +0530] "GET /index.html HTTP/1.1" 200 23061 "-" "Mozilla/5.0 (compatible; GurujiBot/1.0; +http://www.guruji.com/en/WebmasterFAQ.html)"

72.30.79.32 - - [27/May/2009:22:11:42 +0530] "GET /what_we_do.html HTTP/1.0" 200 20954 "-" "Mozilla/5.0 (compatible; Yahoo! Slurp/3.0; http://help.yahoo.com/help/us/ysearch/slurp)"

Figure 2: Sample Web logs of www.protechsc.net

The Filtering Process as follows:

Step 1:Select the logs which don't contain Robots.txt and request of image files.

Step2: Group by IP address of the logs

Step3: Codify the requested page with following information

Step 4: Store only IP address, visited page-id into database and make use of it for the polynomial preprocessing.

These steps are pictorially presented in figure 3.

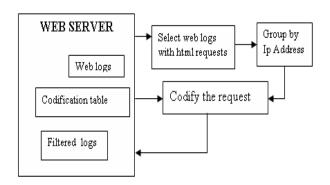


Figure 3. Filtering the Logs

TABLE II - CODIFIED WEBPAGE DETAILS OF A USER
a =[1, 2, 3, 4, 5, 6, 7, 8,13, 0, 0, 0; %1
1, 2, 4, 5,14, 9,10,11,13, 8, 3, 0; %2
1, 2, 4, 5,14,10,11, 5,15,10,13, 4; %3
5,15, 9,10,11,12,13, 6, 8, 7, 0, 0; %4
5, 7, 8, 3, 4, 6, 0, 0, 0, 0, 0, 0; %5
5,16, 9,10,11,12, 3, 7, 8, 0, 0, 0; %6
5,17,10,11, 3, 6, 8, 1, 0, 0, 0, 0; %7
5,26, 9,10,11,12, 5,18, 9,10,11,12; %8
2, 3, 5,27,10,11,12, 6, 8, 0, 0, 0; %9
2, 4, 7, 5,19,10,11,12,13, 0, 0, 0; %10
2, 6, 5, 3, 4, 0, 0, 0, 0, 0, 0, 0; %11
3, 4, 5, 6, 7, 8, 5,32, 9,10,11, 0; %12
3, 5, 8, 0, 0, 0, 0, 0, 0, 0, 0, 0; %13
3, 7, 5,45, 9,10,11,12, 0, 0, 0, 0; %14
1, 4, 5, 7, 8, 3, 6, 5,38, 9,10,11; %15
4, 6, 3, 5,41, 9,10,11,12,13, 8, 1; %16
4, 5,18, 9,10,11,12, 2, 0, 0, 0, 0; %17
4, 6, 5, 8, 3, 5, 0, 0, 0, 0, 0, 0; %18
6, 3, 5, 7, 8, 1, 0, 0, 0, 0, 0, 0; %19
6, 7, 4, 3, 5,22, 5,34, 5,17, 9, 8; %20
1, 4, 5,22, 9,10,11, 3, 0, 0, 0, 0; %21
2, 4, 8, 5,29, 5,32, 5,40, 9,10,11; %22
3, 6, 7, 5,14, 5,16, 5, 4, 0, 0, 0; %23
7, 8, 3, 4, 5, 0, 0, 0, 0, 0, 0, 0; %24
1, 3, 4, 5,14, 9,10,11,12, 2,13, 0] %25

B. Polynomial Interpolation

Polynomial interpolation is the interpolation of a given navigation patterns by a polynomial set obtained by outer product the given navigation sequence. Polynomial interpolation forms the basis for comparing information between two points. The pre-processing generates a polynomial decision boundary. The pre-processing of the input vector is done as follows:

Let X represents the normalized input vector,

$$X = \{X_i\}; i=1,...nf,$$
 (1)

Where X_i is the feature of the input vector

nf is the number of features (nf = 11).

An outer product matrix X_{op} of the original input vector is formed, and it is given by:

Using the X_{op} matrix, the following polynomials are generated:

(i) Product of inputs (NL1)

it is denoted by:

 $\sum w_{ij}x_i$ ($i\neq j$) = Off-diagonal elements of the outer product matrix. (3)

The pre-processed input vector is a 55-dimensional vector.

ii) Quadratic terms (NL2)

It is denoted by: $\Sigma w_{ij}x_i^2 = Diagonal$ elements of the outer product matrix. (4)

The pre-processed input vector is a 11-dimensional vector.

iii) A combination of product of inputs and quadratic terms (NL3)

It is denoted by:

 $\sum w_{ij}x^{i}(i\neq j) + \sum w_{ij}x_{i}^{2} = Diagonal elements and Off-diagonal elements of the outer product matrix. (5)$

The pre-processed input vector is a 66 dimensional vector.

iv) Linear plus NL1 (NL4)

The pre-processed input vector is a 66 dimensional vector. (6)

v) Linear plus NL2 (NL5)

The pre-processed input vector is a 22-dimensional vector. (7)

vi) Linear plus NL3 (NL6)

The pre-processed input vector is a 55-dimensional vector. (8)

In the above polynomials such as NL4, NL5 and NL6 vector, the term 'linear' represents the normalized input pattern without pre-processing. When the training of the network is done with a fixed pre-processing of the input vector, the number of iterations required is less than that required for the training of the network without pre-processing of the input vector to reach the desired MSE. The combinations of different pre-processing methods with different synaptic weight update algorithms are shown in Table III. BPA weight update algorithms have been used with fixed pre-processed input vectors for learning.

C. Back Propagation Algorithm

A neural network is constructed by highly interconnected processing units (nodes or neurons) which perform simple mathematical operations. Neural networks are characterized by their topologies, weight vectors and activation function which are used in the hidden layers and output layer. The

topology refers to the number of hidden layers and connection between nodes in the hidden layers. The activation functions that can be used are sigmoid, hyperbolic tangent and sine. The network models can be static or dynamic. Static networks include single layer perceptrons and multilayer perceptrons. A perceptron or adaptive linear element (ADALINE) refers to a computing unit. This forms the basic building block for neural networks. The input to a perceptron is the summation of input pattern vectors by weight vectors. In most of the applications one hidden layer is sufficient. The activation function which is used to train the Artificial Neural Network is the sigmoid function.

1) Training

- 1. Read log files and filter it
- 2. Separate the data into inputs and target
- 3. Preprocess the data to any NL
- 4. Calculate Principal Component Vector by Z=Z*ZT (9)
 Where Z denotes the cleaned logs
- 5. Train the BPA.
- 5.a Forward Propagation
- (i) The weights of the network are initialized.
- (ii) The inputs and outputs of a pattern are presented to the network
- (iii) The output of each node in the successive layers is calculated.

O (output of a node) =
$$1/(1+\exp(-\sum W_{ii} X_i))$$
 (10)

(iv) The error of a pattern is calculated

$$E(p) = (1/2) \sum (d(p)-o(p))^2$$
 (11)

- 5.b Backward Propagation
- (i) The error for the nodes in the output layer is calculated $\delta(\text{output layer}) = o(1-o)(d-o)$ (12)
- (ii) Weights between output layer and hidden layer are updated.

$$W(n+1) = W(n) + \eta \delta(\text{output layer}) \text{ o(hidden layer)}$$
 (13)

- (iii) The error for the nodes in the hidden layer is calculated. $\delta(\text{hidden layer}) = o(1-o) \sum \delta(\text{output layer})$ W (updated weights between hidden and output layer) (14)
- (iv) The weights between hidden and input layer are updated

$$W(n+1) = W(n) + \eta \delta(hidden layer) o(input layer)$$
 (15)

The above steps complete one weight updating. Second pattern is presented and the above steps are followed for the second weight updating. When all the training patterns are presented, a cycle of iteration or epoch is completed. The errors of all the training patterns are calculated and displayed on the monitor as the mean squared error (MSE).

- 2) Testing
- 1. Read filtered logs and separate into inputs and target
- 2. Preprocess the data with a polynomial function
- 3. Process with final weights of BPA
- 4. Generate the suggestions from the output layer

5. Present the suggestions through templates

IV. RESULTS AND DISCUSSION

Figure 4 presents the mean squared error and classification performance of BPA without preprocessing the input vectors. Fig. 5 to Fig. 10 presents the MSE and classification performance of BPA with preprocessed input vectors. The computational effort, Mean squared error, the iterations required for various algorithm are presented in Table IV. From the Table III, it can be noted that, the algorithm with (BPA +NL2) requires less number of computational effort to achieve minimum 80% classification.

V. CONCLUSION

In this work, a preprocessing approach has been implemented for ANN to learn the web usage mining. The number of arithmetic operations required to train the network with a pre- processed input vector is more, indicating that the computational effort is more. The number of iterations required is less than that required for the vector without pre-processing. The classification performance after preprocessing is more than that of the network trained without pre-processing. The proposed method has to be tried with different types of web sites.

REFERENCES

- [1] Chau, M.; Chen, H., Incorporating Web Analysis Into Neural Networks: An Example in Hopfield Net Searching, IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, Volume 37, Issue 3, May 2007 Page(s): 352 – 358
- [2] David Martens, Bart Baesens, and Tony Van Gestel, Decompositional Rule Extraction from Support Vector Machines by Active Learning, IEEE Transactions On Knowledge And Data Engineering, Vol. 21, No. 2, pp.178 – 191, February 2009
- [3] Dilhan Perera, Judy Kay, Irena Koprinska, Kalina Yacef, and Osmar R. Zaý ane, Clustering and Sequential Pattern Mining of Online Collaborative Learning Data, IEEE Transactions On Knowledge And Data Engineering, Vol. 21, No. 6, pp.759-772 June 2009
- [4] Edmond H.Wu, Michael K.Ng, Joshua Z. Huang, A Data Warehousing and Data Mining Framework for Web usage Management, Communication in Information And Systems Vol. 4, No.4 pp 301-324, 2004
- [5] Guang-Bin Huang, Qin-Yu, Chee-Kheong Siew, Real-Time Learning Capability of Neural Networks, IEEE Transactions on Neural Networks, Vol.17, No.4 July 2006, pp 863-878.
- [6] Hai H. Dam, Hussein A. Abbass, Chris Lokan, and Xin Yao, Neural-Based Learning Classifier Systems, IEEE Transactions On Knowledge And Data Engineering, Vol. 20, No. 1, pp. 26 – 39, January 2008
- [7] Hungjun Lu, Rudy Setiono , Huan Liu , Effective Data mining using neural networks, IEEE Transactions on knowledge and data engineering Vol.8 No.6 December 1996 pp 957-961, 1996.
- [8] James Caverlee, Ling Liu, QA-Pagelet: Data Preparation Techniques for Large-Scale Data Analysis of the Deep Web, IEEE Transactions on knowledge and data engineering Vol.17 No.9 September 2005 pp 1247-1261, 2005

- [9] Lotfi Ben Romdhane, A Softmin-Based Neural Model for Casual Reasoning, IEEE Transactions on Neural Networks, Vol.17, No.3 May 2006, pp 732-744
- [10] Magdalini Eirinaki and Michalis Vazirgiannis, Web Mining For Web Personalization, ACM Transactions on Internet Technology, Vol 3. No.1, February 2003 Pages 1-27
- [11] Mankuan Vai, Zhimin Xu, Representing Knowledge by Neural Networks for qualitative Analysis and Reasoning, IEEE Transactions on knowledge and data engineering Vol.7 No.5 October 1995, pp 683-690
- [12] C. Porcel , E. Herrera-Viedma , Dealing with incomplete information in a fuzzy linguistic recommender system to disseminate information in university digital libraries, ELSEVIER, Knowledge-Based Systems 23 (2010), pp. 40–47
- [13] Ranieri Baraglia and Fabrizio Silvestri, Dynamic Personalization of Web sites without user intervention, Communications of the ACM February 2007, Vol.50 No.2 pp 63-67
- [14] Sankar K.Pal, Pabitra Mirta, Web Mining in Soft Computing Framework: Relevance, State of the Art and Future Directions, IEEE Transactions on Neural Networks, Vol. 13, No. 5 September 2002 pp 1163-1176
- [15] Tianyi Jiang and Alexander Tuzhilin, Improving Personalization Solutions through Optimal Segmentation of Customer Bases, IEEE Transactions On Knowledge And Data Engineering, Vol. 21, No. 3, pp.305-320, March 2009.
- [16] Vir V.Phoha, S.Sitharama iyengar, Rajgopal Kannan, Faster Web Page Allocation with Neural Networks, IEEE Internet Computing November-December 2002. pp 18-26
- [17] Xiaozhe Wang, Ajith Abraham, Kate A. Smith, Intelligent web traffic mining and analysis, Journal of Network and Computer Applications 28 (2005) 147-165, ELSEVIER
- [18] Yu-Hui Tao a, Tzung-Pei Hong b, Yu-Ming Su c , Web usage mining with intentional browsing data, ELSEVIER Expert Systems with Applications, pp.1893–1904. Available online at www.sciencedirect.com 2008, www.elsevier.com/locate/eswa
- [19] Zhicheng Dou, Ruihua Song, Ji-Rong Wen, and Xiaojie Yuan, Evaluating the Effectiveness of Personalized Web Search, IEEE Transactions On Knowledge And Data Engineering, Vol. 21, No. 8, pp.1178 – 1190, August 2009
- [20] Zi Lu Ruiling Han, Jie Duan, Analyzing the effect of website information flow on realistic human flow using intelligent decision models, ELSEVIER, Knowledge-Based Systems 23 (2010), pp. 40–47

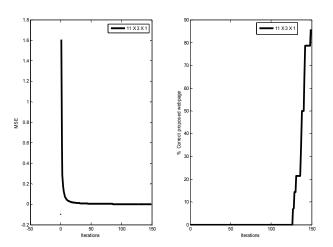


Figure 4. MSE and percentage of correct proposed webpage using BPA without preprocessing the input vector (Table II)

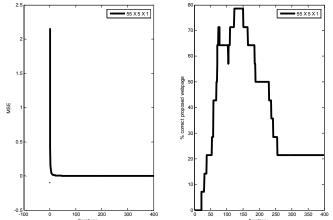


Figure 5 MSE and percentage of correct proposed webpage using (BPA+NL1) with preprocessing the input vector (Table II)

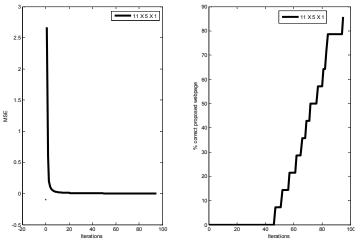


Figure 6. MSE and percentage of correct proposed webpage using (BPA+NL2) with preprocessing the input vector (Table II)

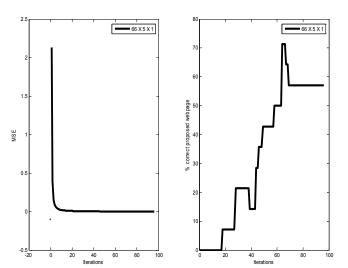


Figure .7 MSE and percentage of correct proposed webpage using (BPA+NL3) with preprocessing the input vector (Table II)

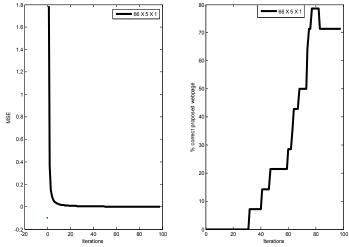


Figure.8 MSE and percentage of correct proposed webpage using (BPA+NL4) with preprocessing the input vector (Table II)

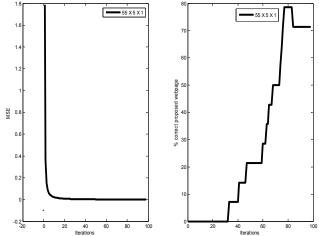


Figure 10 MSE and percentage of correct proposed webpage using (BPA+NL6) with preprocessing the input vector (Table II)

AUTHORS PROFILE

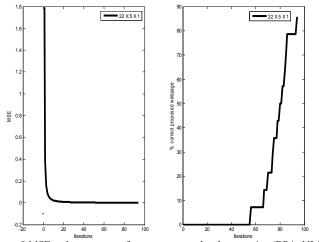


Figure 9.MSE and percentage of correct proposed webpage using (BPA+NL5) with preprocessing the input vector (Table II)

S.S. Scie Uni her Wo

S.Santhi received her B.Sc and M.Sc degrees in Computer Science from University of Madras and Alagappa University in 1997 and 2000 respectively. She completed her M.Phil in Computer Science from Mother Teresa Womens' University in 2003. Her areas of research includes Data Mining and Neural Networks.



Dr. S. Purushothaman is working as professor in Sun College of Engineering, Nagerkoil,India. He received his Ph.D from IIT Madras. His area of research includes Artificial Neural Networks, Image Processing and signal processing. He published more than 50 research papers in national and international journals.

Efficient Probabilistic Classification Methods for NIDS

S.M.Aqil Burney

M.Sadiq Ali Khan

Mr.Jawed Naseem

Department of Computer Science University of Karachi, Karachi-Pakistan Department of Computer Science University of Karachi, Karchi-Pakistan Principal Scientific Officer-PARC

Abstract: As technology improve, attackers are trying to get access of the network system resources by so many means, open loop holes in the network allow them to penetrate in the network more easily. Various approaches are tried for classification of attacks. In this paper we have compared two methods Naïve Bayes and Junction Tree Algorithm on reduced set of features by improving the performance as compared to full data set. For feature reduction PCA is used that helped in proposing a new method for efficient classification. We proposed a Bayesian network-based model with reduced set of features for Intrusion Detection. Our proposed method generates a less false positive rate that increase the detection efficiency by reducing the workload and that increase the overall performance of an IDS. We also investigated that whether conditional independence really effect on the attacks/ threats detection.

Keywords-Network Intrusion Detection System(NIDS); Bayesain Networks; Junction Tree Algorithm

I. Introduction

Network Security whether in a commercial organization or in a critically important research network, is a major issue of concern with the increasing use of web even the personal information in under threat. Efficient network intrusion detection system is only solution to such threats [4].

IDS is a monitoring system of networks to control / avoid / secure the networks from cyber terrorist or it is the process of examing the events occurring in a network or computer system and detecting the signs of incidents which are the threats of computer security policies. Network system monitored by the IDS for detection of any rules violation. Having such violation in the system, efficient IDS generates notification by means of an alarm generation that alert the administrator to put some steps/major according to such vulnerabilities. Common intrusion attacks are classified based on various features/ parameter. KDD-99 data set usually used for investigating the nature of attack. The data set has 41 features listed. Information value of these features and interdependence among them is an interest of investigation. How much reduction in features can be made without reducing the efficiency of classification algorithm and whether interdependency really contributes to detection efficiency? We are tried to find the answers of such kind of questions in this paper. PCA is an effective data dimension reduction technique. Similarly Naïve Bayes' classifier and Bayesian Network both use probabilistic

approach for determination of attack probability. Naïve Bayes' classifiers assume conditional independence while Bayesian network consider assumes conditional dependence. Two methods can be used to compare whether conditional independency or interdependency really contribute to probability of attack. In the next section we discussed some related works which are already proposed, in section 3 we discussed the two methods of classification, in section 4 the methodology is mentioned and finally in section 5 results and discussions are presented.

II. BACKGROUND

For intrusion most network based systems become the target to the hacker, so building efficient IDS is the main task now a day [4]. Intrusion based systems needs a component that generates an alerts on the basis of rule set, to detect the malicious activity correctly it is necessary to manage the alerts correctly [1]. Data Mining approaches are being applied by researchers for the attacks detection in their Intrusion Detection Systems[2]..Probabilistic approaches for reducing the false alarm rate are proposed for example, see [3]. The enormous amount of network data traffic is accumulated each day. Numbers of data mining approaches are used for collecting knowledge domain for intrusion detection which includes clustering, association rules and classification [12]. Data analysis supports by data mining techniques and now it becomes one of the important features/component in intrusion based system. The main concern of using data mining techniques in attacks detection system to differentiate between normal packet vs abnormal. For applying data mining in intrusion detection we need a data set and a classification model. That classification model may be Bayesian Network, neural network, rule based decision tree based and other soft computing techniques as Support Vector Machines(SVM) [10,11]. Intrusion Detection System is now becomes the necessicity for an organizational security system with its credibility that may depend upon the data mining techniques.

2.1 Clustering

The process of labeling data and arranging it in groups is called clustering. By grouping we basically improve the performance of different classifiers used. The genuine cluster contains data corresponding to single category [5]. The data set belongs to the cluster is modeled with respect to them exciting

features. You may define the term clustering in such a way that it refers as unsupervised machine learning mechanism for patterns matching in unlabeled data with numerous aspects.

2.2 Classification

In classification we break the data sets into different classes and it is much less exploratory than clustering. By means of classification we need to classify data into set of classes normal /not normal and to sub classify into different types. Naïve Bayes' used as a classification algorithm in this research by which data classification for intrusion detection be achieved. Due to the collection of huge amount of data traffic needed classification is less famous [6].

III. CLASSIFICATION METHODS

3.1 Naïve Bayes Classifier

Naïve Bayes classifier is an effective technique for classification of data. The technique is particularly useful for large data dimension. The Naïve Bayes is a special case of Bayes theoram which presuppose independence in data attributes [7]. Even though Naïve Bayes assumes data independence, its performance is efficient and at par with other techniques assuming data conditionality. Naïve Bayes classifier can manage continuous or categorical data. Let for a set of given variable $X=\{x_1,x_2,....x_n\}$ with possible outcomes $O=\{o_1,o_2,....o_n\}$. The posterior probability of the dependent variable is obtained by Bayes rule.

$$P(A_i | A) = \frac{P(A_i).P(A|A_i)}{\sum_{j=1}^{N} P(A_j).P(A|A_j)}$$

$$P(O_j | x_1, x_2,, x_n) * P(x_1, x_2,, x_n)O_j P(O_j)$$

We can obtain a new case with X with a class label $\mathbf{O_j}$ have highest posterior probability as

$$P(C_j|X) * P(C_j) \prod_{k=1}^{d} d P(X_k|C_j)$$

The efficiency of Naive Bayes classifier lies in the fact that it converts multi dimensionality of data to one dimensional density estimation. The occupations of evidence do not affect the posterior probability so generally classification task is efficient. The same is proved in this study also when Naive

Bayes classifier is compared with Junction Tree algorithm. For modeling Naive Bayes classifier several distribution including normal gamma or Poisson density function can be employed.

3.2 Junction Tree Algorithm

Its a graphical method of belief updation or probabilistic reasoning. For Probabilistic reasoning, we are using Bayesian Networks and Decision Graphs (BNDG) for which details can be found in [9]. The basic concept in junction tree is clustering of predicted attributes [8]. In belief updation instead of approximating joint probability distribution of all targeted variable (cliques) cluster attributes are formed and potential of clusters are used to approximate probability. So basically junction tree is the graphical representation of potential cluster nodes or cliques and a suitable algorithm to update this potential. Junction tree algorithm involve several steps as moralizing the graph, triangulation junction tree formulation, assigning probabilities to cliques, message passing and reading cliques marginal potentials from junction tree.

Using Junction tree algorithm requires that directed graph

is changed to undirected graph to ensure uniform application process is called moralization which involve adding edges between parents and dropping the direction let $\vec{G} = (N_{\vec{G}}, E_{\vec{G}})$

be a directed graph to be changed into undirected graph $G\left(N_G,E_G\right)$ so infect two new sets along with EG required to be added i.e.

$$E_{\vec{G} \to G}$$
 and $E_{\vec{G} \to G}$

The set can be defined as

$$E_{\mathcal{C} \to \mathcal{C}} := \left\{ (Y_i, Y_j) \in \mathbb{N}^2 : \exists Y_k \in \mathbb{N} \to : (Y_i, Y_k) \in E_{\mathcal{C}} \text{ and } (X_i, X_k) \in E_{\mathcal{C}} \right\}$$

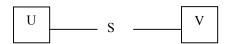
$$E_{\vec{G} \leftrightarrow \vec{G}} := \left\lfloor \left(Y_j, Y_k\right) e N^2 : \left(Y_j, Y_k\right) \notin E_G but\left(Y_k, Y_j\right) \subset E \right\rfloor$$

In moralization $E_{\vec{c}}UE_{\vec{c}\to c}UE_{\vec{c}\to \vec{c}}$ is obtained and new undirected moralized graph is given as

$$G := \left(N_{\vec{G}}, E_{\vec{G}} \right) U E_{\vec{G} \to \vec{G}} \ U E_{\vec{G}}$$

Junction tree is formed after moralization which is basically hyper graphs of cliques if cliques of undirected graph G is given by C(G) than junction tree with a unique property that intersections of any two nodes is contained in every node in the unique path joining the nodes.

Let consider a cluster representation having to neighbor cluster U and V sharing a variable S in common



The aim of JTA is to modify potential in such a way that the distribution of P (V) is obtained by modified potential $\Psi(V)$. In such case probability of S can be given as

$$P(S) = \sum \Psi(V)$$

$$\Psi(U)$$

$$\Psi(S)$$

$$\Psi(V)$$
Similarly

$$P(S) = \sum \Psi(U)$$

Let $\Psi(S)$ represent modified potential so $\Psi(S) = P(S)$, so now if potential of let say $\Psi(V)$ is delayed as result of new evidence f the potential of both $\Psi(S)$ & $\Psi(U)$ can be updated realizing the equivalence

$$\Psi(U) = P(S) = \Psi(V)$$

Belief updation in junction tree is carried out through message passing let U and V are two adjacent node with separator S. so the task is to absorb V and W through S. potential $\Psi(W)$ and $\Psi(S)$ with condition

$$\sum \Psi^*(W) = \Psi^*(S) = \sum \Psi^*(V)$$

In absorption $\Psi^*(S)$ and $\Psi^*(W)$ are replaced as under $\Psi^*(S) = \sum \Psi(V)$

$$\Psi^*(W) = \Psi(W) \frac{\Psi(S)}{\Psi(S)}$$

In this way belief of the whole network is updated through message passing.

IV. METHODOLOGY

KDD'99 data set of intrusion detection was used. PCA technique was used and 14 features were selected on the basis of analysis. Selection of data set for training and testing plays a vital role in accuracy of prediction. In intrusion detection frequency of some attacks are very large as compare to others. To ensure inclusion of all attacks type in learning stratified random sample were drawn relative to proportion of each attack type. This produces better result as compare to simple

random sampling. For Naive Bayes classification two data sets (stratified sample of equal size of 10000) were used for learning and testing using software BN *classifier*. In junction tree algorithm structure learning is carried out by drawing a random sample of 5000 from KDD data sets using *netica*. Then five data sets each of size 1000 are selected through simple random sample, data set is used for learning and drawing junction tree. Data set 2 to 5 were used for testing belief update learned by junction tree.

V. RESULTS & DISCUSSION

The 41 features of KDD'99 data set were reduced to 14 features. The PCA identified 12 major components having Eigen values greater than and around more than 80% variability of data explained by these features while 98% variability can be explained 24 components.

The difference of variability between 24 and 14 features selection is only 18% but computational cost highly increased if 24 parameters are selected, so optimize the processing speed 14 has been selected. It is evident from the graph mentioned above that first 24 components represent 98.866% data and 14 components explained 80% variability which is quite sufficient, and work was carried out on these components only, neglecting the other components which seem less worthy. Besides this, structure learning also support selection of 14 features. The Bayesian network model shown in Figure 2 represents interdependence among various attributes. It is evident that mainly two factors as **count & src_byte** are effected by various features and in turn these two ultimately affect the attack types. The KDD'99 data set classification list 18 attack types however **normal & neptune** are more frequent.

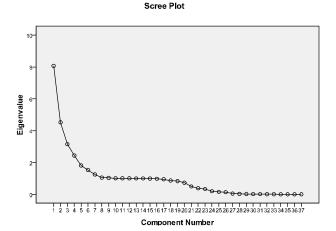


Figure 1: Scree Plot of attributes.

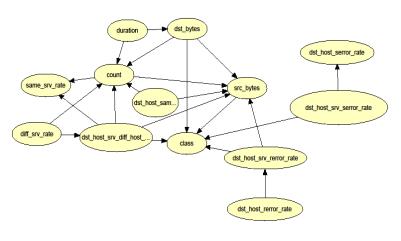


Figure: 2 Bayesian Network Model Intrusion Detection System

BN classification also supports the importance of these two type *normal* (0.527) and *neptune* (0.399) in Table 1. The probability of features buffer overflow, *imap and multihop* are less than 0.001% and that of *ftp_write*, *guess_password* and *load_module* are close to 0. It suggests that this classification can be merged.

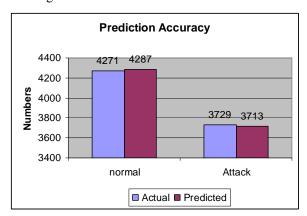


Figure3: Prediction accuracy using BN Classifier

Figure 4 shows majors attacks category predictions. DoS attacks are 99.86% detected while probe attacks about 75% detected.

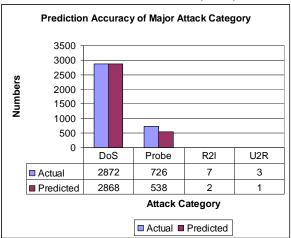


Figure 4: Prediction Accuracy of Major Attacks

BN classifier learned more effectively the attack which is more frequent. In case of identify normal attacks it showed error rate of **0.8%** only and identification of most frequent attack *neptune* is 6.8% refers in table 1.

TABLE 1 ACCURACY OF CLASSIFICATION(BAYESIAN CLASSIFIER)

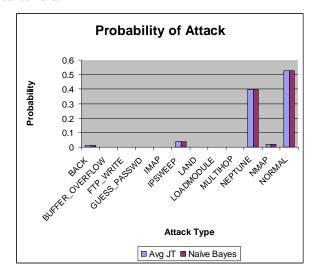
Class	Actual	Predicted	Diff	Error %
back	62	62	0	0
buffer_overflow	2	0	2	100
guess_passwd	3	0	3	100
imap	2	0	2	100
ipsweep	225	284	-59	-26.2
multihop	1	0	1	100
neptune	2630	2587	43	1.6
nmap	96	35	61	63.5
normal	4271	4287	-16	-0.37
phf	1	0	1	100
pod	12	0	12	100
portsweep	186	219	-33	-17.7
rootkit	1	0	1	100
satan	219	273	-54	-24.6
smurf	168	180	-12	-7.1
teardrop	60	39	21	35
warezclient	57	34	23	40.35
warezmaster	4	0	4	100
Total	8000	8000		

TABLE 2. PROBABILITY OF ATTACK(AVERAGE)

Class	Junction	Naïve Bayes	Diff
	Tree	Classifier	
back	0.0102	0.0086	0.0016
buffer_overflow	0.0008	0.001	-0.0002
imap	0.0006	0.0005	0.0001
ipsweep	0.0368	0.0368	0
multihop	0.0002	0	0.0002
neptune	0.3992	0.3936	0.0056
nmap	0.0176	0.0147	0.0029
normal	0.527	0.5432	-0.0162
Total	1	1	

Vol. 8, No. 8, November 2010

Using junction tree algorithm accuracy of identification is utmost 98%. Junction tree also identified *neptune* as most frequent attack. Probability identified of various attacks is depicted in table 2. It is evident that estimation of probability almost equal. This has been statistically compared that there is no significance difference between two methods. Frequencies of remaining attacks are very small and their probability almost near to zero.



VI. CONCLUSION & FUTURE RECOMMENDATIONS

Despite the fact that Naïve Bayes classifiers assume conditional independence and junction tree algorithm parameter interdependence, even though Naïve Bayes and junction tree classifiers are almost equally effective. It is recommended that only those attacks should be considered which are more frequents in order to achieve better performance. It is also found that in selection of learning and testing data set appropriate sampling techniques are utilized for better result prediction.

REFERENCES

- [1] Moon Sun Shin, Eun Hee Kim, and Keun Ho Ryu, "False Alarm classification model for network-based IDS"; Springer-verlag berlin Heidelberg, LNCS 3177, pp. 259–265, 2004.
- [2] M.J.Lee, M.S.Shin, H.S.Moon, "Design and implementation of alert analyzer with data mining engine. Proc. IDEAL '03, Hongkong, 2003.
- [3] A.Valdes and K. Skinner, "Probabilistic alert correlation"; 4th international symposium on Recent Advances in ID, RAID, 54-68, 2003.
- [4] S.M.Aqil Burney and M.Sadiq Ali Khan , "Network Usage Security Policies for Academic Institutions", International Journal of Computer Applications, October Issue, Published By Foundation of Computer Science, 2010.
- [5] Anoop Singhal and Sushil Jajodia, "Data warehousing and data mining techniques for intrusion detection systems", Distributed and Parallel Databases Volume 20, Number 2, 149-166, DOI: 10.1007/s10619-006-9496-5,2006.

- [6] Tasleem Mustafa, Ahmed Mateen, Ahsan Raza Sattar, Nauman ul Haq and M. Yahya Saeed, "Forensic Data Security for Intrusions", European Journal of Scientific Research ISSN 1450-216X Vol.39 No.2 (2010), pp.296-308,2010.
- [7] Karl Friston, Carlton Chu, Jnaina Mourao, Oliver Hulme, Geriant Rees, Will Penny and John Ashburner, "Bayesian decoding of brain images", Elsevier NeuroImage Volume 39, Issue 1, 1, Pages 181-205, January 2008.
- [8] Jaydip Sen, "An agent-based intrusion detection system for local area networks", IJCNIS, Vol. 2, No. 2, August 2010.
- [9] F.V.Jensen and T.S.nielsen, "Bayesian Networks and Decision Graphs" Springer.Berlin Heidelberg, New York, 2007.
- [10] C.Cortes and V. Vapnik," Support Vector Networks". Machine Learning, 20, 1995, pp. 273-297,1995.
- [11] Jungtaek Seo," An Attack Classification Mechanism Based on Multiple Support Vector Machines", LNCS 4706, Part II, pp. 94–103, Springer-Verlag Berlin Heidelberg, ICCSA 2007.
- [12] Hebah H. O. Nasereddin, "Stream Data Mining", International Journal of Web Applications, Volume 1 Number 4 December 2009.

AUTHORS PROFILE



Dr.S.M.Aqil Burney is the Meritorious Professor and approved Supervisor in Computer Science and Statistics by the Higher Education Commission, Govt of Pakistan. He is also the Director & Chairman of Computer Science Department, University of Karachi. Additionally he is also a Director of Main Communication Network University of Karachi. He is also member of various higher academic boards of different universities of Pakistan. His research interest includes AI, Soft Computing, Neural Network, Fuzzy Logic, Data Mining, Statistics, Simulation and Stochastic Modeling of Mobile Communication system and Networks, Network Security and MIS in health services. Dr.Burney is also referee of various journals and conferences proceedings, nationally & internationally. He is member of IEEE(USA), ACM(USA) and



M.Sadiq Ali Khan received his BS & MS Degree in Computer Engineering from SSUET in 1998 and 2003 respectively. Since 2003 he is serving Computer Science Department, University of Karachi as an Assistant Professor. He has about 12 years of teaching experience and his research areas includes Data Communication & Networks, Network Security, Cryptography issues and Security in Wireless Networks. He is member of CSI, PEC and NSP.



Jawed Naseem is Principal Scientific Officer in Pakistan Agricultural Research Council. He has M.Sc(Statistics) and MCS from University of Karachi, currently doing MS (Computer Science) from University of Karachi. His research interest are data modeling, Information Management & Security and Decision Support System particularly in agricultural research. He has been a team member in development of several regional(SAARC) level agricultural databases.

A Survey on Digital Image Enhancement Techniques

V.Saradhadevi¹, Dr.V.Sundaram²

¹Research scholar, Karpagam University, ² Director of MCA, karpagam Engineering College, Coimbatore, India.

Abstract---Image enhancement is one of the major research fields in image processing. In many applications such as medical application, military application, media etc., the image enhancement plays an important role. There are many techniques proposed by different authors in order to remove the noise from the image and produce the clear visual of the image. Also, there are many filters and image smoothing methods available. All these available techniques are designed for particular kind of noises. Recently, neural networks turn to be a very effective tool to support the image enhancement. Neural network is applied in image enhancement because it provides many advantages over the other techniques. Also, neural network can be suitable for removal of all kinds of noises based on its training data. This paper provides survey about some of the techniques applied for image enhancement. This survey deals with the several existing methods for image enhancement using neural networks.

Keywords--- Image Enhancement, Image Denoising, Neural Network, Image Filter, Image Restoration.

I. INTRODUCTION

The intention of image enhancement is to improve the interpretability or perception of data in images for human visual or to provide better input for other automated image processing techniques.

Image enhancement methods can be broadly divided into two categories:

- Spatial domain methods, which involves direct operation on image pixels, and
- Frequency domain methods, which involves Fourier transform of an image for its operation.

Regrettably, there is no general theory for determining what good image enhancement is when it comes to human perception. If it looks good, it is good! However, when image enhancement methods are used as pre-processing tools for other image processing methods, then quantitative measures can decide which techniques are most suitable.

Image Restoration is the technique of retaining the original image from the degraded image given the knowledge of the degrading factors. There are a variety of reasons that could cause degradation of an image and image restoration is one of the key fields in today's Digital Image Processing due to its wide area of applications. Commonly occurring degradations include blurring, motion and noise. Blurring can be caused when object in the image is outside the camera's depth of field sometime during the exposure, whereas motion blur can be caused when an object moves relative to the camera during an exposure. The general model for image degradation phenomenon is given as y = Hf + n, where y is the observed blurred and noisy image, f is the original image, n is additive random noise and H is the blurring operator. The main objective is to estimate the original image from the observed degraded image. Whatever the degraded process, image distortions can fall into two categories, namely, spatially invariant or space invariant and spatially variant or space variant. In a space invariant distortion all pixels have suffered the same form of distortion. This is generally caused by problems with the imaging system such as distortions in optical system, global lack of focus, or camera motion. In a space variant distortion, the degradation suffered by a pixel in the image depends upon its location in the image. This is because of internal factors, such as distortions in the optical system, or by external factors, such as object motion. This survey provides many techniques available for image enhancement.

II. LITERATURE SURVEY

Uma *et al.*, [1] proposed a Morphological Neural Network for color image restoration. This paper considers the problem of color image restoration degraded by a blur function and corrupted by random noise. A new approach developed by multilayer morphological (MLM) neural network is presented, which uses highly nonlinear morphological neuron for image processing to get a high quality restored color image. In this paper color images are considered into RGB distribution. Then each subspace can be considered as a gray image space and is processed by morphological way used in gray images. This method is advantageous because of its low computational overhead, improved performance in terms of signal to noise ratio with less number of neurons.

Gallo *et al.*, [2] presented an adaptive image restoration using local neural approach. This work aims at usage of neural learning for defining and experimentally evaluating an iterative strategy for blind image restoration in the presence of

blur and noise. A salient aspect of the solution is the local estimation of the restored image based on gradient descent strategies able to estimate both the blurring function and the regularized terms adaptively. As an alternative of explicitly defining the values of local regularization parameters through predefined functions, an adaptive learning approach is proposed.

The various restoration techniques used currently can be broadly viewed under two categories, namely, the transform related techniques and the algebraic restoration techniques [3]. The transform related techniques involve analyzing the degraded image after an appropriate transform has been applied. The two popular transform related techniques are inverse filtering and Kalman filtering [4]. Inverse filtering produces a perfect restoration in the absence of noise, but the presence of noise has very bad effects. The Kalman filter approach can be applied to non stationary image but it is computationally very intensive.

Algebraic techniques attempt to find a direct solution to the distortion by matrix inversion techniques, or techniques involving an iterative method to minimize a degradation measure. The two popular algebraic techniques available are pseudo inverse filtering and constrained image restoration. The pseudo inverse spatial image restoration techniques attempt to restore an image by considering the vector space model of the image degradation and attempting to restore the image in this vector space domain. This method does not consider the effects of noise in the calculations of the pseudo inverse and so is sensitive to noise in the image. This involves determining an approximation to the inverse of the matrix blurring operator which is multiplied with the column scanned image vector to produce the degraded image. Blur matrices are very large and it is not computationally feasible to invert them. Constrained restoration techniques are often based on Wiener estimation and regression techniques. One of the major drawbacks in most of the image restoration algorithms is the computational complexity, so various simplifying assumptions have been made to obtain computationally feasible algorithms.

Motivated by the biological neural network in which the processing power lies in a large number of neurons linked with synaptic weights, artificial neural network models attempt to achieve a good performance via dense interconnection of simple computational elements. Neural network models have great potential in areas where high computation rates are required and the current best systems are far from equaling human performance. Restoration of a high quality image from a degraded recording is a good application area of neural nets. Joon *et al.*, [5] proposed a Modified Hopfield neural network model for solving the restoration problem which improves upon the algorithm proposed by Zhou *et al.* [6].

Osman *et al.*, [7] gives an image enhancement using bright and dark stretching techniques for tissue based tuberculosis bacilli detection. This paper proposes two methods for color image enhancement; bright stretching and dark stretching algorithms. Both techniques are well known to create good

image enhancement for gray scale images. But, the current study has adapted these techniques to be used for color images. Even though the adapted image processing method is quite simple, the results signify that these methods may have some potential to be used for improving the quality of Ziehl Neelsen slide images. The experimental result illustrates that both methods proposed by the author can improve the image contrast and enhances the image quality when compared to its conventional techniques.

Pattern learning based image restoration using neural networks is put forth by Dillon *et al.*, [8]. The author illustrate a generic pattern learning based image restoration scheme for degraded digital images, where a feed-forward neural network is employed for implementation of the proposed techniques. The methodology reported here can be applied in several circumstances, for instance, quality enhancement as a post-processing of image compression schemes, blur image restoration and noise image filter, provided that the training data set is comprised of patterns rich enough for supervised learning. This paper focuses on the problem of coded image restoration. The key points addressed in this work are

- The use of edge data extracted from source image as a priori knowledge in the regularization function to get better details and reduce the ringing artifact of the coded images.
- The theoretic basis of the pattern learning-based technique using implicit function theorem.
- Subjective quality improvement with the use of an image similarity for training neural networks
- Empirical studies with contrast to the set partitioning in hierarchical tree (SPIHT) method.

The main advantages of this model-based neural image restoration approach comprise strong robustness with respect to transmission noise and the parallel processing for real-time applications.

Reeves [9] described fast and direct image restoration with edge-preserving regularization. In several applications, fast restorations are required to keep up with the frame rate. FFTbased restoration affords a fast implementation, but it does so at the expense of assuming that the degree of regularization is constant over the image. Unfortunately, this hypothesis can generate significant ringing artifacts in the presence of edges as well as edges that are blurrier than necessary. Shift-variant regularization affords a way to vary the roughness penalty as a function of spatial coordinates. Virtually all edge-preserving regularization techniques exploit this concept. However, this technique destroys the structure that makes the use of the FFT possible, since the deblurring operation is no longer shiftinvariant. Thus, the restoration techniques available for this problem no longer have the computational efficiency of the FFT. The author proposes a new restoration method for the shift-variant regularization approach that can be implemented in a fast and flexible manner. This paper decomposes the restoration into a sum of two independent restorations. One restoration yields an image that comes directly from an FFTbased approach. This image is a shift-invariant restoration consisting of usual artifacts. The other restoration involves a set of unknowns whose number equals the number of pixels with a local smoothing penalty significantly different from the typical value in the image. This restoration represents the artifact correction image. By summing the two, the artifacts are canceled. Since the second restoration has a significantly reduced set of unknowns, it can be calculated very efficiently even though no circular convolution structure exists.

Noise-refined image enhancement using multi-objective optimization is illustrated by Peng et al., [10]. This paper presents a novel scheme for the enhancement of images using stochastic resonance (SR) noise. In this scheme, a suitable dose of noise is added to the lower quality images such that the performance of a suboptimal image enhancer is improved without altering its parameters. In this paper, image enhancement is modeled as a constrained multi-objective optimization (MOO) problem, with similarity and some desired image enhancement characteristic being the two objective functions. The principle of SR noise-refined image enhancement is analyzed, and an image enhancement system is developed. A genetic algorithm-based MOO technique is employed to find the optimum parameters of the SR noise distribution. In addition, a novel image quality evaluation metric based on human visual system (HVS) is developed as one of the objective functions to guide the MOO search procedure.

Lu et al., [11] proposed an image noise reduction technique based on the fuzzy rules. Considering the image as nonstationary signal, an image noise reduction method based on the fuzzy rules is proposed. This image processing system (IPS) is recognized as a time-variant system in which the system parameters change continuously based on the local characteristics of the images. For the purpose of noise reduction, Gaussian noise is considered here. The fuzzy rules are implemented to consider the unstableness and uncertainty of signals. The nonlinear function indicating the fuzzy rulebased IPS depends on the rules concerning the local characteristics of the input, on the membership functions, and on the used defuzzification method. For making the system performance as high as possible, these factors must be agreed to be the most appropriate ones. In this paper a technique for designing the optimum nonlinear function directly from the local characteristics of training data is presented. Here the rules, the membership functions, and the technique of defuzzification are not essential to be known. The design of these factors is concerned in the design of the membership function, thus attaining the optimum nonlinear function is sufficient for designing the IPS. The only thing required to do is to choose what sort of the local characteristics of the image should be applied to the rule-based system. Computer simulations illustrate that the proposed technique gives better results in comparison with that of the weighted averaging filter and median filter.

An adaptive fuzzy image enhancement algorithm for local regions is given by Yan *et al.*, [12]. To overcome the drawbacks of low speed and losing image information in fuzzy image enhancement algorithms, a novel fuzzy enhancement operator with close-character and transplantable-character is

proposed in this paper. The approached operator utilizes the gradient operator to create the image enhancement processing focus on the interested regions, and the OTSU operator to automatically select the best threshold value, which can realize a novel adaptive fuzzy image enhancement algorithm for local regions. Through the experimentations of the asphalt pavement crack image detection system, the experimental results specify that the novel algorithm can not only attain better processing effects and higher processing speed than now-available fuzzy image enhancement algorithms, but also possess the property of high practicability and generality.

Faouzi *et al.*, [13] provides a directional-rational approach for color image enhancement. In this, the author presents an unsharp masking-based approach for noise smoothing and edge enhancing in multichannel images. The structure presented by author is similar to the conventional unsharp masking structure, however, the enhancement is allowed only in the direction of maximal change and the enhancement parameter is computed as a nonlinear function of the rate of change. This scheme improves the true details, limits the overshoot near sharp edges and attenuates noise in flat areas. In addition the use of the control function eliminates the need for the subjective coefficient λ used in the conventional unsharp masking method.

The noise reduction based on fuzzy image filtering is put forth by Dimitri et al., [14]. A new fuzzy filter is provided for the noise reduction of images corrupted with additive noise. The filter involves two stages. The initial stage calculates a fuzzy derivative for eight different directions. The next stage uses these fuzzy derivatives to carry out fuzzy smoothing by weighting the contributions of neighboring pixel values. Both these stages are dependent on fuzzy rules which make use of membership functions. The filter can be implemented iteratively to effectively decrease heavy noise. Especially, the shape of the membership functions is adapted according to the remaining noise level after each iteration; making use of the distribution of the homogeneity in the image. A statistical technique for the noise distribution can be included to relate the homogeneity to the adaptation scheme of the membership functions.

Gacsadi *et al.*, [15] makes use of cellular neural network for the purpose of image enhancement. This technique takes both the denoising and the increase of the contrast into consideration. Due to whole parallel processing, computing-time reduction is achieved. In the enhancement process by usage of nonlinear and feedback template local and also regional properties will be taken into consideration due to the propagation of the effect between the neighbors. Considerable computing power is required to solve the image processing task described by variational computing. The Cellular Neural Networks (CNN) proved to be very useful regarding real-time image processing. The reduction of computing time, due to parallel processing, can be obtained only if the processing algorithm can be implemented on a CNNUC or by using emulated digital CNN-UM implemented on FPGAs.

The traditional analog architectures of CNN-UC are superior in terms of processing speed and power dissipation. However, these implementations have a restricted applicability due to their limited features regarding accuracy, flexibility, small number of cells, their high cost and the lengthy period needed for the development of such a chip. On the other hand, while software solutions are extremely flexible, they are sometimes inefficient because of limited and low processing speed. Today, the version of CNN digital emulator implemented FPGA is a solution that achieved a compromise between speed and accuracy, but ensure repeatability, reproducibility, flexibility, possibility for CNN implementation even for complex processes of processing and easy interfacing with digital systems. In this sense, the CNN digital emulator implemented FPGA maximizes, for a concrete application, the performances of the overall CNN processing.

There are two basic approaches to image denoising [16] - spatial filtering methods and transform domain filtering methods.

Spatial Filtering

A conventional way to remove noise from image data is to employ spatial filters. Spatial filters can be further classified into non-linear and linear filters.

i. Non-Linear Filters

With non-linear filters, the noise is removed without any attempts to explicitly identify it. Spatial filters utilize a low pass filtering on groups of pixels with the assumption that the noise occupies the higher region of frequency spectrum. Generally spatial filters remove noise to a reasonable extent but at the cost of blurring images which in turn makes the edges in pictures invisible. In recent years, a variety of nonlinear median type filters such as weighted median, rank conditioned rank selection, and relaxed median have been developed to overcome this drawback.

ii. Linear Filters

A mean filter is the optimal linear filter for Gaussian noise in the sense of mean square error. Linear filters also tend to blur sharp edges, destroy lines and other fine image details, and perform poorly in the presence of signal-dependent noise. The wiener filtering method requires the information about the spectra of the noise and the original signal and it works well only if the underlying signal is smooth. Wiener method implements spatial smoothing and its model complexity control correspond to choosing the window size.

Sarode et al., [17] proposed the color image enhancement with the help of fuzzy system. This technique involves the use of knowledge-base (fuzzy expert) systems that are capable of mimicking the behavior of a human expert. Fuzzy technique of knowing severity of tumor is essential to determine if there is a need for the biopsy and it gives to user a clear idea of spread and severity level of tumor. Fuzzy based improvement of color feature of tumor is an application of fuzzy in the area of color feature extraction for enhancement of a peculiar feature. It has been determined that RGB color model is not

appropriate for enhancement because the color components are not decoupled. Alternatively, in HSV color model, hue (H), the color content, is separate from saturation (S), which can be used to dilute the color content and V, the intensity of the color content. By conserving H, and modifying only S and V, it is likely to enhance color image. Therefore, it is required to convert RGB into HSV for the purpose. A Gaussian type membership function is utilized to model S and V property of the image. This membership function utilizes only one fuzzifier and is evaluated by maximizing fuzzy contrast.

Muthu Selvi *et al.*, [18] put forth a hybrid image enhancement technique for noisy dim images using curvelet and morphology techniques. The noisy dim images degrade the image quality. The denoising method using curvelet transform outperforms than wavelet transform. The noisy dim image is made noise free with the help of curvelet transform to the dim image for avoiding the over illumination and under illumination problems. Next the dim image is enhanced using the morphological transformations. Closing by reconstruction is implemented to identify the background of the dim image. The experimental result shows that the morphological restoration filter with closing by reconstruction produces better result than opening by reconstruction.

Hassan *et al.*, [19] presented a contrast enhancement technique for dark blurred images. The chief goal presented by the author is to produce a contrast enhancement technique to recover an image within a given area, from a blurred and darkness specimen, also improve visual quality of it. This method consists of two steps unsharp masking step and contrast enhancement step. The unsharp masking step is applied to the image to sharpen edges and bring out hidden details. On the contrary enhancement step 3x3 slider map window was applied to the image to determine if the corresponding pixel will be remapped or not. The new value of remapped pixel obtained is based on a sigmoid map function. Good and satisfying results were obtained by experimentation on this technique.

Gilbao et al., [20] uses the complex diffusion processes for image enhancement and denoising. The linear and nonlinear scale spaces, obtained by the inherently real-valued diffusion equation, are generalized to complex diffusion processes, by incorporating the free Schrodinger equation. A basic solution for the linear case of the complex diffusion equation is developed. Investigation of its performance shows that the generalized diffusion process combines properties of both forward and inverse diffusion. It verifies that the imaginary part is a smoothed second derivative, scaled by time, when the complex diffusion coefficient approaches the real axis. Based on this observation, the authors develop two examples of nonlinear complex processes, useful in image processing: a regularized shock filter for image enhancement and a ramp preserving denoising process.

Image denoising using non-negative sparse coding shrinkage algorithm is given by Shang *et al.*, [21]. The author proposes a new method for denoising natural images using this extended non-negative sparse coding (NNSC) neural network shrinkage

algorithm, which is self-adaptive to the statistic property of natural images. The fundamental principle of denoising using NNSC shrinkage is similar to that using standard sparse shrinkage and wavelet soft threshold. Using test images corrupted by additive Gaussian noise, this paper evaluated the method across a range of noise levels. This method utilized the normalized mean squared error as a measure of the quality of denoising images and the signal to noise rate (SNR) value as an evaluative feature of different denoising approaches. The experimental result shows that the NNSC shrinkage certainly is effective in image denoising. Otherwise, the author also compares the effectiveness of the NNSC shrinkage with sparse coding shrinkage and wavelet soft threshold method. The simulative tests show that this denoising method outperforms any other of the two kinds of denoising approaches.

Gupta et al., [22] designed a FIR filter for image restoration using principal component neural network. The neural network can be applied in many image denoising applications because of its inherent characteristics such as nonlinear mapping and self-adaptiveness. The design of filters widely depends on the a-priori knowledge about the type of noise. Because of this, standard filters are application and image specific. Extensively used filtering algorithms reduce noisy artifacts by smoothing. Though, this operation normally results in smoothing of the edges as well. Alternatively, sharpening filters enhance the high frequency details making the image non-smooth. An integrated general technique to design a finite impulse response filter based on principal component neural network (PCNN) is provided in this study for image filtering, optimized in the sense of visual inspection and error metric. This technique utilizes the inter-pixel correlation by iteratively updating the filter coefficients using PCNN. This technique performs optimal smoothing of the noisy image by preserving high and low frequency features. Experimental results state that this filter is robust under various noise distributions. Additionally, the number of unknown parameters is very few and most of these parameters are adaptively obtained from the processed image.

Image denoising based on combined neural networks filter is proposed Junhong *et al.*,[23]. A new image restoration technique based on combined neural networks Alter is proposed by the author. This integrated neural networks Alter is posed by a BPNN Alter and an image data fusion system based on self-organizing mapping neural networks. And this technique can use the corrupted image itself as training data to avoid the problem of how to choose the training data, which is most of the other neural networks denoising methods have to face, by using the distributed character of WGN. Experiment results show that this method can denoise the noises effectively.

Gacsadi et al., [24] describes PDE-based medical images denoising using Cellular Neural Networks. The author presents the medical image denoising by using cellular neural networks (CNN), based on the variational model of Chan and Esedoglu. By comparatively examining the proposed method and other CNN methods that uses variational computation, the

proposed method offering the best efficiency in terms of image denoising and edge preservation.

Zhang et al., [25] presented image denoising using a neural network based non-linear filter in wavelet domain. Images are often distorted as a result of various factors that can occur during acquisition and transmission processes. Image denoising is intended at removing or reducing noise, so that a good-quality image can be obtained for various applications. The author presents a neural network based denoising method implemented in the wavelet transform domain. A noisy image is first wavelet transformed into four subbands, and then a trained layered neural network is applied to each subband to generate noise-removed wavelet coefficients from their noisy ones. The denoised image is then obtained through the inverse transform on the noise-removed wavelet coefficients. Compared with other techniques performed in the wavelet domain, it requires no a priori knowledge about the noise and needs only one level of signal decomposition to obtain very good denoising results.

III. CONCLUSION

This survey discusses about several existing image enhancement methods. All those methods discussed have their own advantages and disadvantages. This survey helps in choosing the better suitable image enhancement scheme for particular kind of noise in the image. Also, the filtering algorithms used for removing the noise from the image are presented in this thesis. This survey explains about the importance of neural network for image enhancement as the neural networks have the advantages such as nonlinear mapping and self-adaptiveness. To overcome the demerit of the techniques discussed in this survey, Adaptive Neuro-Fuzzy Interference Systems (ANFIS) can be used for image enhancement as it combines the advantages of Artificial Neural Networks (ANN) and Fuzzy Interference System (FIS).

REFERENCES

- [1] S. Uma and S. Annadurai, "Color Image Restoration Using Morphological Neural Network", ICGST.
- [2] I. Gallo, E. Binaghi and A. Macchi, "Adaptive Image Restoration using a Local Neural Approach".
- [3] H. C. Andrews & B. R. Hunt, "Digital Image restoration", Englewood cliffs, NJ, Prentice Hall, 1977.
- [4] Rafael C. Gonzalez and Richard E. Woods, "Digital image processing", 2nd edition, Addison-Wesely, 2004.
- [5] Joon K. Paik and Aggelos K. Katsaggelos, "Image restoration using a modified Hopfield Network", IEEE Transactions on image processing, Vol 1, No.1, pp. 49-63, January 1992.
- [6] Y.T.Zhou, R. Chellappa and B.K. Jenkins, "Image restoration using a neural network", IEEE Trans. Acoust., Speech, Signal Processing, Vol, ASSP-36, pp 1141-1151, July 1988.
- [7] M.K. Osman, M.Y. Mashor and H. Jaafar2Colour, "Image Enhancement using Bright and Dark Stretching Techniques for Tissue based Tuberculosis Bacilli Detection", Proceedings of the International Conference on Man-Machine Systems (ICoMMS), 2009.
- [8] Dianhui Wang Dillon and T. Chang, "Pattern learning based image restoration using neural networks", Proceedings of the International Joint Conference on Neural Network, 2002.

- [9] S.J. Reeves, "Fast and direct image restoration with edge-preserving regularization", IEEE Digital Signal Processing Workshop, 2002.
- [10] Renbin Peng, Hao Chen and Varshney, "Noise-refined image enhancement using multi-objective optimization", 44th Annual Conference on Information Sciences and Systems (CISS), 2010.
- [11] Ruihua Lu and Li Deng, "An Image Noise Reduction Technique Based on the Fuzzy Rules".
- [12] Yan Maode, Bo Shaobo, Li Xue and He Yuyao, "An Adaptive Fuzzy Image Enhancement Algorithm for Local Regions", Chinese Control Conference, 2007.
- [13] Faouzi Alaya Cheikh and Moncef Gabbouj, "Directional-Rational Approach for Color Image Enhancement".
- [14] Dimitri Van De Ville, Mike Nachtegael, Dietrich Van der Weken, Etienne E. Kerre, Wilfried Philips, and Ignace Lemahieu, "Noise Reduction by Fuzzy Image Filtering", IEEE transactions on fuzzy systems, vol. 11, no. 4, august 2003.
- [15] A. Gacsadi, V. Tiponut, E. Gergely, I. Gavrilut, "Variational Based Image Enhancement Method by using Cellular Neural Networks", Proceedings of the 13th WSEAS International Conference on SYSTEMS
- [16] Mukesh C. Motwani, Mukesh C. Gadiya, Rakhi C. Motwani and Frederick C. Harris, Jr., "Survey of Image Denoising Techniques".
- [17] Milindkumar V. Sarode, S.A.Ladhake and Prashant R. Deshmukh, "Fuzzy system for color image enhancement", World Academy of Science, Engineering and Technology, 2008.
- [18] Muthu Selvi, Roselin and Kavitha, "A Hybrid Image Enhancement Technique for Noisy Dim Images Using Curvelet and Morphology", International Journal of Engineering Science and Technology Vol. 2(7), 2010.
- [19] Naglaa Yehya Hassan and Norio Aakamatsu, "Contrast Enhancement Technique of Dark Blurred Image", IJCSNS International Journal of Computer Science and Network Security, Vol.6, No.2A, February 2006.
- [20] Gilboa G, Sochen N and Zeevi Y.Y, "Image enhancement and denoising by complex diffusion processes", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004.
- [21] Li Shang and Deshuang Huang, "Image denoising using non-negative sparse coding shrinkage algorithm", IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005.
- [22] Gupta.P.K and Kanhirodan.R, "Design of a FIR Filter for Image Restoration using Principal Component Neural Networks", IEEE International Conference on Industrial Technology, 2006.
- [23] Junhong Chen and Qinyu Zhang, "Image Denoising Based on Combined Neural Networks Filter", International Conference on Information Engineering and Computer Science, 2009.
- [24] Gacsadi.A, Grava.C, Straciuc.O and Gavrilut.I, "PDE-based medical images denoising using Cellular Neural Networks", International Symposium on Signals, Circuits and Systems, 2009.
- [25] Zhang.S and Salari.E, "Image denoising using a neural network based non-linear filter in wavelet domain", IEEE International Conference on Acoustics, Speech and Signal Processing, 2005.

A Survey on Designing Metrics suite to Asses the Quality of Ontology

K.R Uthayan Department of Information Technology, SSN College of Engineering Chennai, India G.S.Anandha Mala, Professor & Head, Department of Computer Science & Engineering St.Joseph's College of Engineering, Chennai, India

Abstract---With the persistent growth of the World Wide Web, the difficulty is increased in the retrieval of relevant information for a user's query. Present search engines offer the user with several web pages, but different levels of relevancy. To overcome this, the Semantic Web has been proposed by various authors to retrieve and utilize additional semantic information from the web. As the Semantic Web adds importance for sharing knowledge on the internet this has guide to the development and publishing of several ontologies in different domains. Using the database terminology, it can be said that the web-ontology of a semantic web system is schema of that system. As web ontology is an integral aspect of semantic web systems, hence, design quality of a semantic web system can be deliberated by measuring the quality of its webontology. This survey focuses on developing good ontologies. This survey draws upon semiotic theory to develop a suite of metrics that assess the syntactic, semantic, pragmatic, and social aspects of ontology quality. This research deliberates about the metrics that may contribute in developing a high quality semantic web system.

Keywords--- Quality Metrics, Web ontology, Semiotic Metrics, Semantic Quality, Domain modularity.

I. INTRODUCTION

CEMANTIC Web is nothing but the extension of the present web in which the web resources are prepared with formal semantics about their interpretation for the machines. These web resources are combined in the form of web information systems, and their formal semantics are usually characterized in the form of web-ontologies. By means of the database terminology, it can be said that the web-ontology of a semantic web system is representation of that system [11]. Design quality of a semantic web system can be calculated by computing the quality of its web-ontology because web ontology is the integral element of semantic web systems [25]. The main concern is that when the design of a web-ontology is completed, it is suitable time to assess its quality so that in case, the design is of low quality, it can be enhanced before its instantiation. This helps in saving of considerable amount of cost and effort for developing high quality semantic web systems. Metrics are considered as the appropriate tools for estimating quality. This survey focuses on several metrics for web ontology quality evaluation.

II. LITERATURE SURVEY

Ahluwalia *et al.*, [1] presented a Semiotic Metrics Suite for Assessing the Quality of Ontologies. Table 1 shows some of the metrics for quality evaluation [1, 3].

As a decisive construct, overall quality (Q) is a subjective function of its syntactic (S), semantic (E), pragmatic (P), and social (O) qualities [1] (i.e., $Q = b1 \times S + b2 \times E + b3 \times P + b4 \times O$). The addition of weight is equal to 1. In the absence of pre-specified weights, the weights are assigned to be equal.

Syntactic Quality (S) evaluates the quality of the ontology according to the way it is written. Lawfulness is the extent to which an ontology language's rules have been obeyed. Not every ontology editors have error-checking capabilities; however, without correct syntax, the ontology cannot be read and used. Richness is nothing but the proportion of features in the ontology language that have been used in ontology (e.g., whether it includes terms and axioms, or only terms). Richer ontologies are more valuable to the user (e.g., agent).

Semantic Quality (E) estimates the meaning of terms in the ontology library. Three attributes are used here are interpretability, consistency, and clarity. Interpretability deals with the meaning of terms (e.g., classes and properties) in the ontology. In the real world, the knowledge provided by the ontology can map into meaningful concepts. This is accomplished by checking that the words used by the ontology be present in another independent semantic source, such as a domain-specific lexical database or a comprehensive, generic lexical database such as WordNet. Consistency is nothing but whether terms having a consistent meaning in the ontology. For example, if an ontology claims that X is a subclass of Y, and that Y is a property of X, then X and Y have incoherent meanings and are of no semantic value. For example, ontological terms such as IS-A is often used inconsistently. Clarity is the term which determines whether the context of terms is clear. For example, if ontology claims that class "Chair" has the property "Salary," an agent must know that this illustrate academics, not furniture.

Pragmatic Quality (P) deals with the ontology's usefulness for users or their agents, irrespective of syntax or semantics. Three criteria are used for determining P. Accuracy is whether the claims on ontology makes are 'true.' This is very tricky to determine automatically without a learning mechanism or truth maintenance system. Currently, a domain expert evaluates accuracy. The measure of the size of the ontology is

called as Comprehensiveness. Larger ontologies are more probable to be complete representations of their domains, and provide more knowledge to the agent. Relevance indicates whether the ontology satisfies the agent's specific requirements.

TABLE 1: DETERMINATION OF METRIC VALUES

Attributes	Determination	
Overall Quality (Q)	Q = b1.S + b2.E + b3.P + b4.O	
Syntactic Quality (S)	S = bs1.SL + bs2.SR	
Lawfulness (SL)	Let X be total syntactical rules. Let X_b be total breached rules. Let NS be the number of statements in the ontology. Then $SL = X_b / NS$.	
Richness (SR)	Let Y be the total syntactical features available in ontology language. Let Z be the total syntactical features used in this ontology. Then $SR = Z/Y$.	
Semantic Quality (E)	E = be1.EI + be2.EC + be3.EA	
Interpretability (EI)	Let C be the total number of terms used to define classes and properties in ontology. Let W be the number of terms that have a sense listed in WordNet. Then EI = W/C.	
Consistency (EC)	Let I = 0. Let C be the number of classes and properties in ontology. ∀Ci, if meaning in ontology is inconsistent, I+1. Therefore, I = number of terms with inconsistent meaning. Ec = I/C.	
Clarity (EA)	Let Ci = name of class or property in ontology. \forall Ci, count Ai, (the number of word senses for that term in WordNet). Then EA = A/C.	
Pragmatic Quality (P)	P = bp1.PO + bp2.PU + bp3.PR	
Comprehensiveness (PO)	Let C be the total number of classes and properties in ontology. Let V be the average value for C across entire library. Then PO = C/V.	
Accuracy (PU)	Let NS be the number of statements in ontology. Let F be the number of false statements. PU = F/NS. Requires evaluation by domain expert and/or truth maintenance system.	
Relevance (PR)	Let NS be the number of statements in the ontology. Let S be the type of syntax relevant to agent. Let R be the number of statements within NS that use S. PR = R / NS.	
Social Quality (O)	O = bo1.OT + bo2.OH	
Authority (OT)	Let an ontology in the library be OA. Let the set of other ontologies in the library be L. Let the total number of links from ontologies in L to OA be K. Let the average value for K across ontology library be V. Then OT = K/V.	
History (OH)	Let the total number of accesses to an ontology be A. Let the average value for A across ontology library be H. Then OH = A/H.	
Cohesion (Coh)	Coh= SCC Where SCC is separate connected components	
Fullness (F)	$F = \frac{ C_i(I) }{ C_i(I) }$	
Readability (Rd)	Rd = A, A = rdfs: comment + A, A = rdfs: label	

For the purpose of evaluation, it needs some knowledge of the agent's requirements. This metric is coarse as it verifies for the type of information the agent uses by ontology (e.g., property, subclass, etc), rather than the semantics needed for

specific tasks (e.g., the particular subclasses needed to interpret a user's specific query).

Social quality (O) imitates the fact that agents and ontologies exist in communities. The authority of an ontology is nothing but the number of other ontologies that link to it (define their terms using its definitions). More authoritative ontologies indicate that the knowledge they provide is accurate or useful. The history indicates the number of times the ontology is accessed. Ontologies are more dependable when they are with longer histories.

The cohesion (Coh) of a KB is nothing but the number of separate connected components (SCC) of the graph representing the KB.

The fullness (F) of a class C_i is defined as the actual number of instances that belong to the subtree rooted at C_i ($C_i(I)$) compared to the expected number of instances that belong to the subtree rooted at C_i ($C_i(I)$).

The readability (Rd) of a class C_i is defined as the total of the number attributes that are comments and the number of attributes that are labels the class has.

Amjad *et al.*, [2] provided the Web-Ontology Design Quality Metrics. The author proposes design metrics for web-ontology [21] by maintaining certain recommended principles like a metric may reach its highest value for perfect quality for excellent case and vice versa that is it may reach its lowest level for worst case. It is supposed to be monotonic, clear, and intuitive. It must correlate well with human decisions and it should be automated if possible. The proposed metrics may give notification about how much knowledge can be derived from a given webontology; how much it is relevant to a user's specific necessities and how much it is effortless to reuse, manage, trace and adapt. The metrics provided by the author are Knowledge Enriched (KnE), Characteristics Relevancy (ChR) and Domains modularity (DoM).

Knowledge Enriched metric

The reasoning capability of a web-ontology is determined by Knowledge Enriched (KnE) metric, and it is based on two sub-metrics so-called Isolated Axiom Enriched (IAE) metric and Overlapped Axiom Enriched (OAE) metric. There are three parts in this axiom namely, predicate, resource and object. If none of these is similar with any other axiom of identical domain then that axiom is termed as isolated axiom. If the two axioms have some similar parts, it is said to be overlapped. There may be more than a few transitively overlapped axioms in any domain. This metric determines the percentage of IAE and OAE, and if the former is greater than the later one, then the web-ontology can be regarded as less knowledge enriched. IAE is officially defined as the ratio of total number of isolated axioms (tIAs) to the total number of domain axioms (tDAs).

$$IAE = \sum_{i=1}^{n} \frac{tIAs}{tDAs}$$

$$for \ all \ 1 \le i \le n$$
(1)

In the above equation, n is total number of sub-domains of web-ontology. Similarly, the OAE metric is officially defined as ratio of total number of overlapped axioms (tOAs) to the total number of domain axioms. It can be written as follows:

$$OAE = \sum_{i=1}^{n} \frac{tOA_i}{tDAs}$$

$$for \ all \ 1 \le i \le n$$
(2)

In the equation given above, n is total number of sub-domains of web-ontology. Lastly, the KnE metric is the difference of total number of overlapped axioms and the total number of isolated axioms. It may be written as follows:

$$KnE = OAE - IAE$$
 (3)

If the resultant KnE value is positive, then the web-ontology is more knowledge enriched, if it is zero, then the web-ontology is average knowledge enriched, and if it is negative, then the web-ontology is less knowledge enriched.

Characteristics Relevancy metric

Characteristics Relevancy (ChR) metric gives us the suggestion about how much a given web-ontology is close to a user's specific necessities and the degree of reusability of the web-ontology. Formally, it is termed as the ratio of the number of relevant attributes (nRAs) in a class to the total number of attributes (TnAs) of that class. It can be written as follows:

$$ChR = \sum_{i=1}^{n} \frac{nRAs}{TnAs}$$

$$for \ all \ 1 \le i \le n$$
(4)

where n in above equation represents the total number of classes in the provided web-ontology. ChR metric reveals the proportion of relevant attributes in the web-ontology, and this number gives insights how much a web-ontology is relevant.

Domain Modularity metric

Domain modularity (DoM) metric denotes the component-orientation feature of a web-ontology. This metric specifies the grouping of knowledge in different components of web-ontology. The webontology is best manageable, traceable, reusable and adaptable, if it is designed in components (subdomains). Formally, the DoM metric is given as the number of sub-domains (NSD) contained in a webontology. This metric also depends on the coupling and cohesion [25] levels of sub-domains, and it is directly proportional to its cohesion level and inversely proportional to its coupling level.

$$DoM = NSD + \sum_{i=1}^{N} DCoh_i + 1 / \sum_{i=1}^{n} DCoup_i$$
 for all $1 \le i \le n$ (5)

In the above equation, DCoh indicates the level of domain cohesion and DCoup represents the level of coupling among sub-domains of web-ontology domain. DoM metric is a real number indicating the degree of partial reusability of a given web-ontology.

Samir et al., [3] given the OntoQA: Metric-Based Ontology Quality Analysis. The metrics presented can highlight key characteristics of an ontology schema and also its population and facilitate users to make an informed judgment easily. The metrics used by the author here are not 'gold standard' measures of ontologies. Instead, the metrics are projected to estimate several aspects of ontologies and their potential for knowledge representation. Rather than describing ontology as merely effective or ineffective, metrics describe a certain aspect of the ontology because, in most cases, the way the ontology is built is largely dependent on the domain in which it is designed. The metrics defined here are Schema Metrics and Instance Metrics. The following are metrics considered by the author:

The following are some of Schema Metrics:

Relationship Richness: The diversity of relations and placement of relations in the ontology is defined by this metrics. An ontology that has many relations further than class-subclass relations is better than taxonomy with no more than class-subclass relationships. The relationship richness (RR) is defined as the ratio of the number of relationships (P) defined in the schema to the sum of the number of subclasses (SC) plus the number of relationships.

$$RR = \frac{|P|}{|SC| + |P|}$$

Attribute Richness: The attribute richness (AR) is defined as the average number of attributes (slots) per class. It is given as the ratio of number attributes for all classes (att) to the number of classes (C).

$$AR = \frac{|att|}{|C|}$$

Inheritance Richness: The inheritance richness of the schema (IRs) is defined as the average number of subclasses per class. The number of subclasses (C1) for a class Ci is defined as IHC (C1, Ci)l.

$$IR_s = \frac{\sum_{C_i \in C} |H^C(C_1, C_i)|}{|C|}$$

The following are some of Instance Metrics:

Class Richness: The class richness (CR) of a knowledge base is defined as the ratio of the number of classes used in the base (C`) to the number of classes defined in the ontology schema (C).

$$CR = \frac{|C^{\hat{}}|}{|C|}$$

Average Population: Formally, the average population (P) of classes in a knowledge base is defined as the number of instances of the knowledge base (I) to the number of classes defined in the ontology schema (C).

$$P = \frac{|I|}{|C|}$$

Importance: The importance (Imp) of a class Ci is defined as the number of instances that belong to the subtree rooted at Ci in the knowledge base (Ci(I)) compared to the total number of instances in the knowledge base (I).

$$Imp = \frac{|C_i(I)|}{|I|}$$

Werner [4] provided a Realism-Based Metric for Quality Assurance in Ontology Matching. There are three levels introduced to the methodology for the measurement of quality improvements in single ontologies. These levels are:

- Level 1: reality, consisting of both instances and universals and also the various relations that acquire between them;
- Level 2: the cognitive representations of this reality personified in observations and interpretations;
- Level 3: the publicly accessible concretizations of the cognitive representations in representational artifacts of a range of sorts, of which ontologies are examples.

Harith *et al.*, [5] defined the metrics for Ranking Ontologies. In this paper AKTiveRank, a prototype system for ranking ontologies is proposed based on the analysis of their structures. This paper describes the metrics used in the ranking system. The ranking measures used are described below:

Class Match Measure

The Class Match Measure (CMM) is intended to estimate the coverage of ontology for the provided search terms. AKTiveRank looks for classes in every ontology that have labels matching a search term either exactly (class label identical to search term) or partially (class label "contains" the search term).

Density Measure

Density Measure (DEM) is deliberated to approximate the representational-density or information-content of classes and accordingly the level of knowledge detail. DEM considers how well the concept is additionally specified (the number of

subclasses), the number of attributes related with that concept, number of siblings, etc.

Semantic Similarity Measure

Similarity measures have often been used in information retrieval systems to afford enhanced ranking for query results. Ontologies can be analyzed as semantic graphs of concepts and relations, and hence similarity measures can be applied to explore these conceptual graphs. This helps in resolving ambiguities.

Henry [7, 23] described a Measurement Ontology Generalizable for Emerging Domain Applications on the Semantic Web. The semantic Web is considered as the next generation Web of structured data that are automatically shared by software agents, which apply definitions and constraints structured in ontologies to correctly process data from contrasting sources. One aspect needed to develop semantic Web ontologies of emerging domains is creating ontologies of concepts that are common to those domains. These general ontologies can be used as building blocks to develop more domain-specific ontologies. However most measurement ontologies focus on representing units of measurement and quantities, and not on other measurement concepts such as sampling, mean values, and evaluations of quality based on measurements. In this paper, the author elaborates on a measurement ontology that represents all these concepts. This paper presents the generality of the ontology, and describes how it is developed, used for analysis and validated.

Fensel *et al.*, [8] provided OIL (Ontology Interchange Language): an ontology infrastructure for the Semantic Web. Initially, Researchers in artificial intelligence motivate the development of ontologies [14] to facilitate knowledge sharing and reuse. Ontologies [15] play a key role in supporting information exchange across different networks. A prerequisite for such a role lead to the development of a joint standard for specifying and exchanging ontologies. The authors present OIL which satisfies such standards.

Carlos et al., [9] presented an Ontology-based Metrics Computation for Business Process Analysis. Business Process Management (BPM) aims to support the whole life-cycle required to deploy and maintain business processes in organizations. Analyzing business processes have a need of computing metrics that can facilitate determining the health of business activities and thus the whole enterprise. However, the degree of automation currently achieved cannot maintain the level of reactivity and adaptation demanded by businesses. In this paper the author argue and show how the use of Semantic Web technologies can enhance to an important extent the level of automation for analyzing business processes. The author presents a domain-independent ontological framework for Business Process Analysis (BPA) with support for automatically computing metrics. In particular, a set of ontologies for specifying metrics are defined in this paper. The domain-independent metrics computation engine is defined that can interpret and compute them.

Orme *et al.*, [10] described Coupling Metrics for Ontology-Based Systems. XML has grown to be frequent in Internet-based application domains such as business-to-business and business-to-consumer applications. It has moreover produced a basis for service-oriented architectures such as Web services and the Semantic Web, mainly because ontology data employed in the Semantic Web [16, 17] are stored in XML. Measuring system coupling is a generally accepted software engineering practice connected with producing high-quality software products. In many application domains, coupling can be assessed in ontology-based systems before system development by measuring coupling in ontology data. A proposed set of metrics determines coupling of ontology data in ontology-based systems [22] represented in the Web Ontology Language (OWL), a derivative of XML.

Andrew et al., [1] define a semiotic metrics suite for assessing the quality of ontologies. A suite of metrics proposed here is to assess the quality of the ontology. The metrics evaluate the syntactic, semantic, pragmatic, and social aspects of ontology quality according to the semiotic theory. The author operationalizes the metrics and employs them in a prototype tool called the Ontology Auditor. A primary validation of the Ontology Auditor on the DARPA Agent Markup Language (DAML) library of domain ontologies represents that the metrics are feasible and highlights the wide variation in quality between ontologies in the library. The contribution of the research is to afford a theory-based framework that developers can utilize to develop high quality ontologies and that applications can exploit to choose appropriate ontologies for a given task. Zhe et al., [24] provides some Evaluation Metrics for Ontology Complexity and Evolution Analysis.

Ying et al., [12] discusses about semantic web. Presently, computers are shifting from single, isolated devices into door points to a worldwide network of information exchange and business transactions called the World Wide Web (WWW). For this cause, support in data, information, and knowledge exchange has become a key issue in current computer technology. The achievement of the WWW has made it increasingly hard to find, access, present, and maintain the information required by a wide variety of users. In answer to this problem, many new research initiatives and commercial enterprises have been provided to enhance available information with machine processable semantics. This semantic web will offer intelligent access to heterogeneous, distributed information, enabling software products (agents) [20] to intervene between user needs and the information sources available. This paper reviews ongoing research in the area of the semantic web [19], focusing especially on ontology technology.

Anthony et al., [13, 18] put forward the Complexity and coupling metrics for ontology based information. Ontologies are greatly used in bioinformatics and genomics to characterize the structure of living things. This research focuses on complexity metrics for ontologies. These complexity metrics are obtained from semantic relationships in an ontology. These metrics will assist for selecting the best

ontologies in several application areas, including bioinformatics and genomics.

Dimitris N. Kanellopoulos [22] described ODELO: an ontology-driven model for the evaluation of learning ontologies. Trying out or renewing an existing learning ontology [6] and providing evaluation tools to assess its quality are fundamental steps before putting an e-learning system online. Ontology [21, 23] evaluation is a central task and it is typically the output of an automatic process. This paper put forwards an ontology-driven model, called Ontology-Driven model for the Evaluation of Learning Ontologies (ODELO), for the estimation of ontologies representing learning resources with respect to several metrics. Syntax contracts with the proper relations between signs (e.g., words, phrases, sentences) and the construction of new ones. Social metrics imitate the fact that software agents and ontologies coexist and communicate in communities. Ontology cohesion metrics indicates the degree of relatedness of ontology classes. ODELO is a deductive valuation model that identifies the elements of ontological quality for learning ontologies. In this paper the author propose a framework for assessing the quality of learning ontologies that constitute the basis for intelligent educational Adaptive Hypermedia (AH) systems.

III. CONCLUSION

In this survey, the different suite of metrics for evaluating ontologies based upon the semiotic-web is analyzed. It is better to assess the quality of web-ontology after the design is completed. This helps in neglecting the low quality ontology before its development by enhancing those defects in ontology. This helps in saving of considerable amount of cost and effort [2] for developing high quality semantic web systems. Several metrics such as Knowledge Enriched metric, Characteristics Relevancy metric, Domain Modularity metric, Richness, Instance Metrics, Semantic Similarity Measure, Density Measure, etc., are analyzed for assessing the quality of ontology. This survey helps in choosing the best suited metrics for assessing the quality of ontology. The different metrics which involve different criteria for ontology are analyzed in this survey.

REFERENCES

- Punit Ahluwalia, Andrew Burton-Jones, Veda C. Storey and Vijayan Sugumaran "A Semiotic Metrics Suite for Assessing the Quality of Ontologies", journal of Data & Knowledge Engineering,vol 55, pp 84-102, 2005.
- [2] Amjad Farooq, Syed Ahsan, and Abad Shah, "Web-Ontology Design Quality Metrics", Journal of American Science, 2010.
- [3] Samir Tartir, I. Budak Arpinar, Michael Moore, Amit P. Sheth and Boanerges Aleman-Meza, "OntoQA: Metric-Based Ontology Quality Analysis", IEEE Workshop on Knowledge Acquisition from Distributed, Autonomous, Semantically Heterogeneous Data and Knowledge Sources, 2005.
- [4] Werner Ceusters," Towards a Realism-Based Metric for Quality Assurance in Ontology Matching", Proceeding of the 2006 conference on Formal Ontology in Information Systems: Proceedings of the Fourth International Conference (FOIS), pp 321-332, 2006.

- [5] Harith Alani and Christopher Brewster," Metrics for Ranking Ontologies", WWW2006, pp 22–26, 2006.
- [6] Maedche and Steffen Staab"Ontology Learning for the Semantic Web", IEEE Intelligent Systems, Vol 16, pp 72-79, 2001
- [7] Henry M. Kim," A Measurement Ontology Generalizable for Emerging Domain Applications on the Semantic Web", Journal of Database Management, Vol 18, pp 20-42, 2006.
- [8] D. Fensel., F.V. Harmelen, I. Horrocks., D.L. McGuinness, and P.F. Patel-Schneider, "Oil: An ontology infrastructure for the semantic web", IEEE Intelligent Systems, pp 38-45, 2001.
- [9] Carlos Pedrinaci and John Domingue," Ontology-based Metrics Computation for Business Process Analysis".
- [10] Orme, Anthony M., Yao, Haining, and Etzkorn, Letha, "Coupling Metrics for Ontology-Based Systems," IEEE Software, Vol. 23, No. 2, pp. 102-108, 2006.
- [11] Chen, Y., Zhou, L., & Zhang, D. "Ontology-Supported Web Service Composition: An Approach to Service-Oriented Knowledge Management in Corporate Services." Journal of Database Management, 17(1),pp 67-84, 2006.
- [12] Ying Ding, Dieter Fensel, Michel Klein and Borys Omelayenko, "The semantic web: yet another hip?", Data & Knowledge Engineering, Vol.41 n.2-3, pp 205-227, 2002.
- [13] Anthony Mark Orme, Haining Yao and Letha H. Etzkorn, "Complexity metrics for ontology based information", International Journal of Technology Management, 2009.
- [14] Thomas R. Gruber, "A translation approach to portable ontology specifications", Knowledge Acquisition, Vol.5 no.2, pp.199-220, 1993.
- [15] Martin Doerr, Jane Hunter and Carl Lagoze, "Towards a Core Ontology for Information Integration", Journal of Digital Information, Vol 4, No.1, 2003.
- [16] Jeff Heflin, James Hendler, "A Portrait of the Semantic Web in Action", IEEE Intelligent Systems, Vol.16 no.2, pp.54-59, 2001.
- [17] James Hendler, "Agents and the Semantic Web", IEEE Intelligent Systems, Vol.16 n.2, pp.30-37, 2001.
- [18] Anthony M. Orme, Haining Yao, Letha H. Etzkorn, "Coupling Metrics for Ontology-Based Systems," IEEE Software, Vol. 23, no. 2, pp. 102-108, 2006.
- [19] Alexander Maedche, Steffen Staab, "Ontology Learning for the Semantic Web", IEEE Intelligent Systems, Vol.16 no.2, pp.72-79, 2001.
- [20] Larry M. Stephens, Michael N. Huhns, "Consensus Ontologies: Reconciling the Semantics of Web Pages and Agents", IEEE Internet Computing, Vol.5 n.5, pp.92-95, 2001.
- [21] Vijayan Sugumaran , Veda C. Storey, "Ontologies for conceptual modeling: their creation, use, and management, Data & Knowledge Engineering", Vol.42 no.3, pp.251-271, 2002.
- [22] Dimitris N. Kanellopoulos, "ODELO: an ontology-driven model for the evaluation of learning ontologies", International Journal of Learning Technology archive, Vol 4, pp 73-99, 2009.
- [23] Henry M. Kim, Arijit Sengupta and Joerg Evermann, "MOQ: Web services ontologies for QoS and general quality evaluations".
- [24] Zhe Yang, Dalu Zhang and Chuan YE, "Evaluation Metrics for Ontology Complexity and Evolution Analysis", ICEBE '06. IEEE International Conference, 2006.
- [25] Yinglong Maa, Beihong Jinb and Yulin Fengb, "Semantic oriented ontology cohesion metrics for ontology-based systems", Journal of Systems and Software archive, Vol 83, pp 143-152, 2010.

An Anomaly-Based Network Intrusion Detection System Using Fuzzy Logic

R. Shanmugavadivu

Assistant professor, Department of Computer Science PSG College of Arts & Science, Coimbatore-14

Dr.N.Nagarajan

Principal, Coimbatore Institute of Engineering and Information Technology, Coimbatore.

Abstract—IDS which are increasingly a key part of system defense are used to identify abnormal activities in a computer system. In general, the traditional intrusion detection relies on the extensive knowledge of security experts, in particular, on their familiarity with the computer system to be protected. To reduce this dependence, various data-mining and machine learning techniques have been used in the literature. In the proposed system, we have designed fuzzy logic-based system for effectively identifying the intrusion activities within a network. The proposed fuzzy logic-based system can be able to detect an intrusion behavior of the networks since the rule base contains a better set of rules. Here, we have used automated strategy for generation of fuzzy rules, which are obtained from the definite rules using frequent items. The experiments and evaluations of the proposed intrusion detection system are performed with the KDD Cup 99 intrusion detection dataset. The experimental results clearly show that the proposed system achieved higher precision in identifying whether the records are normal or attack one.

Keywords-Intrusion Detection System (IDS); Anomaly based intrusion detection; Fuzzy logic; Rule learning; KDD Cup 99 dataset.

I. INTRODUCTION

Intrusion incidents to computer systems are increasing because of the commercialization of the Internet and local networks [1]. Computer systems are turning out to be more and more susceptible to attack, due to its extended network connectivity. The usual objective of the aforesaid attacks is to undermine the conventional security processes on the systems and perform actions in excess of the attacker's permissions. These actions could encompass reading secure or confidential data or just doing vicious destruction to the system or user files [2]. A system security operator can detect possibly malicious behaviors as they take place by setting up intricate tools, which incessantly monitors and informs activities [22]. Intrusion detection systems are turning out to be progressively significant in maintaining adequate network protection [1, 3, 4, 5]. An intrusion detection system (IDS) watches networked devices and searches for anomalous or malicious behaviors in the patterns of activity in the audit stream [6]. Capability of discriminating between standard and anomalous user behaviors should be present in a good intrusion detection system [7]. This would comprise of any event, state, content, or behavior that is regarded as abnormal by a pre-defined criterion [8].

Intrusion detection has emerged as a significant field of research, because it is not theoretically possible to set up a system with no vulnerabilities [9]. One main confrontation in intrusion detection is that we have to find out the concealed attacks from a large quantity of routine communication activities [10]. Several machine learning (ML) algorithms, for instance Neural Network [11], Support Vector Machine [12], Genetic Algorithm [13], Fuzzy Logic [14], and Data Mining [15] and more have been extensively employed to detect intrusion activities both known and unknown from large quantity of complex and dynamic datasets. Generating rules is vital for IDSs to differentiate standard behaviors from strange behavior by examining the dataset which is a list of tasks created by the operating system that are registered into a file in historical sorted order [16]. Various researches with data mining as the chief constituent has been carried to find out newly encountered intrusions [17]. The analysis of data to determine relationships and discover concealed patterns of data which otherwise would go unobserved is known as data mining. Many researchers have used data mining to focus into the subject of database intrusion detection in databases [18].

According to the detection strategy used, data miningbased intrusion detection systems can be classified into two main categories [23]. They are misuse detection which identifies intrusions using patterns of well known intrusions or weak spots of the system and anomaly detection, which attempts to find out if departure from the recognized standard usage patterns can be flagged as attacks [19]. (a) Misuse **Detection**: On the basis of the impressions of known intrusions and known system weaknesses misuse detection tries to model abnormal activities. (b) Anomaly Detection: Both user and system behavior can be predicted using normal behavior patterns. Anomaly detectors identify possible attack attempts by constructing profiles representing normal usage and then comparing it with current behavior data to find out a likely mismatch [20]. For specified, well-known intrusion excellent detection results are achieved by signature-based methods. But, they cannot find out unfamiliar intrusions though constructed as a least alteration of previously known attacks. Conversely, the capability of discovering intrusion events which are previously unobserved is the main advantage of anomaly-based detection techniques [21].

In the proposed system, we have designed anomaly based intrusion detection using fuzzy logic. The input to the proposed system is KDD Cup 1999 dataset, which is divided into two subsets such as, training dataset and testing dataset. At first, the training dataset is classified into five subsets so that, four types of attacks (DoS (Denial of Service), R2L (Remote to Local), U2R (User to Root), Probe) and normal data are separated. After that, we simply mine the 1-length frequent items from attack data as well as normal data. These mined frequent items are used to find the important attributes of the input dataset and the identified effective attributes are used to generate a set of definite and indefinite rules using deviation method. Then, we generate fuzzy rule in accordance with the definite rule by fuzzifying it in such a way, we obtain a set of fuzzy if-then rules with consequent parts that represent whether it is a normal data or an abnormal data. These rules are given to the fuzzy rule base to effectively learn the fuzzy system. In the testing phase, the test data is matched with fuzzy rules to detect whether the test data is an abnormal data or a normal data.

The rest of the paper is organized as follows: Section II presents literature review of the proposed system and section III describes the detailed analysis of the KDD cup 99 dataset. The proposed intrusion detection system using fuzzy logic is given in section IV. Experimentation and performance analysis of the proposed system is discussed in section V. Finally, the conclusion is given in section VI.

II. REVIEW OF RECENT RESEARCH

Several techniques are available in the literature for detecting the intrusion behavior. In recent times, intrusion detection has received a lot of interest among the researchers since it is widely applied for preserving the security within a network. Here, we present some of the techniques used for intrusion detection.

S. F. Owens and R. R. Levary [24] have stated that intruder detection systems have been commonly constructed using expert system technology. But, Intrusion Detection System (IDS) researchers have been biased in constructing systems that are difficult to handle, lack insightful user interfaces and are inconvenient to use in real-life circumstances. The proposed adaptive expert system has utilized fuzzy sets to find out attacks. The expert system comparatively easy to implement when used with computer system networks has the capability of adjusting to the nature and/or degree of the threat. Experiments with Clips 6.10 have been used to prove the adjusting capability of the system. Alok Sharma et al. [25] have focused on the use of text processing techniques on the system call sequences for intrusion detection. Host-based intrusions have been detected by introducing a kernel based similarity measure. Processes have been classified either as normal or abnormal using the knearest neighbor (kNN) classifier. They have assessed the proposed method on the DARPA-1998 database and compared its operation with other existing methods present in the literature.

Shi-Jinn Horng *et al.* [26] have used a combination of hierarchical clustering algorithm, easy feature selection method, and SVM technique in their proposed SVM-based intrusion detection system. Fewer, abstracted, and higher-qualified training instances that are derived from the KDD Cup 1999 training set has been given to the SVM by the hierarchical clustering algorithm. The simple feature selection method employed for the removal of insignificant features from the training set has enabled the proposed SVM model to achieve more precise classification of the network traffic data. The proposed system has been assessed using the renowned KDD Cup 1999 dataset. Compared to other intrusion detection systems that are based on the same dataset, the proposed method has exhibited superior performance in identifying DoS and Probe attacks and an overall best performance in accuracy.

B. Shanmugam and Norbik Bashah Idris [28] have proposed an advanced fuzzy and data mining methods based hybrid model to find out both misuse and anomaly attacks. Their objective was to decrease the quantity of data kept for processing and also to improve the detection rate of the existing IDS using attribute selection process and data mining technique respectively. A modified version of APRIORI algorithm which is an improved Kuok fuzzy data mining algorithm utilized for implementing fuzzy rules has enabled the generation of if-then rules that show common ways of expressing security attacks. They have achieved faster decision making using mamdani inference mechanism with three variable inputs in the fuzzy inference engine which they have employed. The DARPA 1999 data set has been used to test and benchmark the efficiency of the proposed model. In addition, the test results against the "live" networking environment within the campus have been analyzed.

O. A. Adebayo et al. [29] have presented a method that uses Fuzzy-Bayesian to detect real-time network anomaly attack for discovering malicious activity against computer network. They have established the effectiveness of the method by describing the framework. The overall performance of the intrusion detection system (IDS) based on Bayes has been improved by a combination of fuzzy with Bayesian classifier. In addition, by the experiment carried out on KDD 1999 IDS data set, the practicability of the method has been verified. Abadeh, M.S. and Habibi, J. [27] have proposed a method to develop fuzzy classification rules for intrusion detection use in computer networks. The method of fuzzy rule base system design has been based on the iterative rule learning approach (IRL). Using the evolutionary algorithm to optimize one fuzzy classifier rule at a time, the fuzzy rule base has been created in an incremental fashion. Intrusion detection problem has been used as a high-dimensional classification problem to analyze the functioning of the final fuzzy classification system. Results have demonstrated that the fuzzy rules generated by the proposed algorithm can be utilized to build a reliable intrusion detection system.

Arman Tajbakhsh *et al.* [30] have presented a data mining technique based framework for constructing an IDS. In the framework, Association Based Classification (ABC) has been used by the classification engine which is in fact the central part of the IDS. Fuzzy association rules have been used by the proposed classification to construct classifiers. Some matching

measures have been used to evaluate the consistency of any new sample (which is to be categorized) with various class rulesets and the label of the sample has been declared as the class that is analogous to the best matched ruleset. A method which decreases the items that may be included in extracted rules has also been proposed to reduce the time taken by the rule induction algorithm. The framework has been assessed using KDD-99 dataset. The results have shown that the achieved total detection rate and detection rate of known attacks are large and false positive rate is small, though the results are not bright for unknown attacks.

Zhenwei Yu et al. [31] have presented an automatically tuning intrusion detection system (ATIDS). According to the feedback supplied by the system operator, when false predictions are detected, the proposed system automatically tunes the detection model on-the-fly. The KDDCup'99 intrusion detection dataset has been used to assess the system. In the experimental results, the system has demonstrated a 35% enhancement with regard to misclassification cost compared to a system that is not using the tuning feature. If the model has been tuned using only 10% false predictions still a 30% improvement is achieved by the system. Moreover, the model tuned using only 1.3% of the false predictions have been capable of achieving about 20% improvement provided the tuning is not delayed too long. Building a practical system based on ATIDS has been proved to be feasible by the results of the experiments: Because predictions ascertained to be false have been used for tuning the detection model, system operators can concentrate on confirmation of predictions with low confidence.

III. KDD CUP 99 DATASET

In 1998, DARPA in concert with Lincoln Laboratory at MIT launched the DARPA 1998 dataset for evaluating IDS [36]. The DARPA 1998 dataset contains seven weeks of training and also two weeks of testing data. In total, there are 38 attacks in training data as well as in testing data. The refined version of DARPA dataset which contains only network data (i.e. Tcpdump data) is termed as KDD dataset [37]. The Third International Knowledge Discovery and Data Mining Tools Competition were held in colligation with KDD-99, the Fifth International Conference on Knowledge Discovery and Data Mining. KDD dataset is a dataset employed for this Third International Knowledge Discovery and Data Mining Tools Competition. KDD training dataset consists of relatively 4,900,000 single connection vectors where each single connection vectors consists of 41 features and is marked as either normal or an attack, with exactly one particular attack type [38]. These features had all forms of continuous and symbolic with extensively varying ranges falling in four categories:

- In a connection, the first category consists of the *intrinsic* features which comprises of the fundamental features of each individual TCP connections. Some of the features for each individual TCP connections are duration of the connection, the type of the protocol (TCP, UDP, etc.) and network service (http, telnet, etc.).
- The *content* features suggested by domain knowledge are used to assess the payload of the original TCP packets, such as the number of failed login attempts.
- Within a connection, the *same host* features observe the recognized connections that have the same destination host as present connection in past two seconds and the statistics related to the protocol behavior, service, etc are estimated.
- The *similar same service* features scrutinize the connections that have the same service as the current connection in past two seconds.

A variety of attacks incorporated in the dataset fall into following four major categories: Denial of Service Attacks: A denial of service attack is an attack where the attacker constructs some computing or memory resource fully occupied or unavailable to manage legitimate requirements, or reject legitimate users right to use a machine. User to Root Attacks: User to Root exploits are a category of exploits where the attacker initiate by accessing a normal user account on the system (possibly achieved by tracking down the passwords, a dictionary attack, or social engineering) and take advantage of some susceptibility to achieve root access to the system. Remote to User Attacks: A Remote to User attack takes place when an attacker who has the capability to send packets to a machine over a network but does not have an account on that machine, makes use of some vulnerability to achieve local access as a user of that machine. Probes: Probing is a category of attacks where an attacker examines a network to collect information or discover well-known vulnerabilities. These network investigations are reasonably valuable for an attacker who is staging an attack in future. An attacker who has a record, of which machines and services are accessible on a given network, can make use of this information to look for fragile points.

Table I illustrates a number of attacks falling into four major categories and table II presents a complete listing of a set of features characterized for the connection records.

TABLE I. VARIOUS TYPES OF ATTACKS DESCRIBED IN FOUR MAJOR CATEGORIES

Denial of Service Attacks	Back, land, neptune, pod, smurf, teardrop
User to Root Attacks	Buffer_overflow, loadmodule, perl, rootkit,
Remote to Local Attacks	Ftp_write, guess_passwd, imap, multihop, phf, spy, warezclient, warezmaster
Probes	Satan, ipsweep, nmap, portsweep

Table II. A complete list of features given in KDD cup 99 dataset

Feature index	Feature name	Description	Туре
1	duration	length (number of seconds) of the connection	continuous
2	protocol_type	type of the protocol, e.g. tcp, udp, etc.	symbolic
3	service	network service on the destination, e.g., http, telnet, etc.	symbolic
4	flag	normal or error status of the connection	symbolic
5	src_bytes	number of data bytes from source to destination	continuous
6	dst_bytes	number of data bytes from destination to source	continuous
7	Land	1 if connection is from/to the same host/port; 0 otherwise	symbolic
8	wrong_fragment	number of ``wrong" fragments	continuous
9	urgent	number of urgent packets	Continuous
10	hot	number of ``hot" indicators	Continuous
11	num_failed_logins	number of failed login attempts	Continuous
12	logged_in	1 if successfully logged in; 0 otherwise	Symbolic
13	num_compromised	number of ``compromised" conditions	Continuous
14	root_shell	1 if root shell is obtained; 0 otherwise	Continuous
15	su_attempted	1 if ``su root" command attempted; 0 otherwise	Continuous
16		number of ``root" accesses	Continuous
17	num_root		
	num_file_creations	number of file creation operations	Continuous
18	num_shells	number of shell prompts	Continuous
19	num_access_files	number of operations on access control files	Continuous
20	num_outbound_cmds	number of outbound commands in an ftp session	Continuous
21	is_hot_login	1 if the login belongs to the ``hot" list; 0 otherwise	Symbolic
22	is_guest_login	1 if the login is a ``guest" login; 0 otherwise	Symbolic
23	count	number of connections to the same host as the current connection in the past two seconds	continuous
24	srv_count	number of connections to the same service as the current connection in the past two seconds	Continuous
25	serror_rate	% of connections that have ``SYN" errors	continuous
26	srv_serror_rate	% of connections that have ``SYN" errors	Continuous
27	rerror_rate	% of connections that have ``REJ" errors	Continuous
28	srv_rerror_rate	% of connections that have ``REJ" errors	Continuous
29	same_srv_rate	% of connections to the same service	Continuous
30	diff_srv_rate	% of connections to different services	Continuous
31	srv_diff_host_rate	% of connections to different hosts	Continuous
32	dst host count	count for destination host	continuous
33	dst_host_srv_count	srv_count for destination host	continuous
34	dst_host_same_srv_rat	same_srv_rate for destination host	continuous
35	e dst_host_diff_srv_rate	diff_srv_rate for destination host	continuous
36	dst_host_same_src_po rt_rate	same_src_port_rate for destination host	continuous
37	dst_host_srv_diff_host rate	diff_host_rate for destination host	continuous
38	dst_host_serror_rate	serror_rate for destination host	continuous
39	dst_host_srv_serror_ra te	srv_serror_rate for destination host	continuous
40	dst_host_rerror_rate	rerror_rate for destination host	continuous
41	dst_host_srv_rerror_ra te	srv_serror_rate for destination host	continuous

IV. AN ANOMALY-BASED NETWORK INTRUSION DETECTION SYSTEM USING FUZZY LOGIC

Presently, it is unfeasible for several computer systems to affirm security to network intrusions with computers increasingly getting connected to public accessible networks (e.g., the Internet). In view of the fact that there is no ideal solution to avoid intrusions from event, it is very significant to detect them at the initial moment of happening and take necessary actions for reducing the likely damage [32]. One approach to handle suspicious behaviors inside a network is an intrusion detection system (IDS). For intrusion detection, a wide variety of techniques have been applied specifically, data mining techniques, artificial intelligence technique and soft computing techniques. Most of the data mining techniques like association rule mining, clustering and classification have been applied on intrusion detection, where classification and pattern mining is an important technique. Similar way, AI techniques such as decision trees, neural networks and fuzzy logic are applied for detecting suspicious activities in a network, in which fuzzy based system provides significant advantages over other AI techniques.

Recently, several researchers focused on fuzzy rule learning for effective intrusion detection using data mining techniques. By taking into consideration these motivational thoughts, we have developed a fuzzy rule based system in detecting the attacks. This system, anomaly-based intrusion detection makes use of effective rules identified in accordance with the designed strategy, which is obtained by mining the data effectively. The fuzzy rules generated from the proposed strategy can be able to provide better classification rate in detecting the intrusion behavior. Even though signature-based systems provide good detection results for specified and familiar attacks, the foremost advantage of anomaly-based detection techniques is their ability to detect formerly unseen and unfamiliar intrusion occurrences. On the other hand and in spite of the expected erroneousness in recognized signature specifications, the rate of false positives in anomaly-based systems is generally higher than in signature based ones [21]. The different steps involved in the proposed system for anomaly-based intrusion detection (shown in figure 1) are described as follows:

- (1) Classification of training data
- (2) Strategy for generation of fuzzy rules
- (3) Fuzzy decision module
- (4) Finding an appropriate classification for a test input

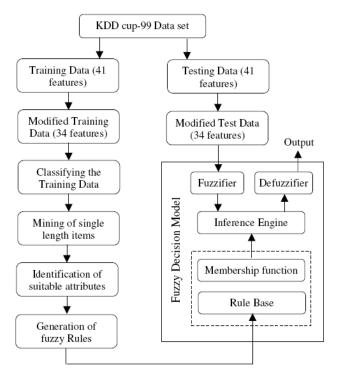


Figure 1. The overall steps of the proposed intrusion detection system

(1) Classification of Training Data

The first component of the proposed system is of classifying the input data into multiple classes by taking in mind the different attacks involved in the intrusion detection dataset. The dataset we have taken for analyzing the intrusion detection behavior using the proposed system is KDD-Cup 1999 data. The detailed analysis of KDD-Cup 1999 data is given in section III. Based on the analysis, the KDD-Cup 1999 data contains four types of attacks and normal behavior data with 41 attributes that have both continuous and symbolic attributes. The proposed system is designed only for the continuous attributes because the major attributes in KDD-Cup 1999 data are continuous in nature. Therefore, we have taken only the continuous attributes for instance, 34 attributes from the input dataset by removing discrete attributes. Then, the dataset (D) is divided into five subsets of classes based on the class label prescribed in the dataset $D = \{D_i : 1 \le i \le 5\}$. The class label describes several attacks, which comes under four major attacks (Denial of Service, Remote to Local, U2R and Probe) along with normal data. The five subsets of data are then used for generating a better set of fuzzy rules automatically so that the fuzzy system can learn the rules effectively.

(2) Strategy For Generation of Fuzzy Rules

This section describes the designed strategy for automatic generation of fuzzy rules to provide effective learning. In general, the fuzzy rules given to the fuzzy system is done manually or by experts, who are given the rules by analyzing intrusion behavior. But, in our case, it is very difficult to generate fuzzy rules manually due to the fact that the input data is huge and also having more attributes. But, a few of researches are available in the literature for automatically

identifying of fuzzy rules in recent times. Motivated by this fact, we make use of mining methods to identify a better set of rules. Here, definite rules obtained from the single length frequent items are used to provide the proper learning of fuzzy system. The process of fuzzy generation is given in the following sub-section.

(a) Mining of single length frequent items

At first, frequent items (attributes) are discovered from both classes of input data and by using these frequent items, the significant attributes are identified for the input KDD-cup 99 dataset. In general, frequent itemset are mined using various conventional mining algorithms, such as Apriori [35] and FP-Growth [40]. These algorithms are suitable to mine frequent itemset with varying length only for the binary database, which contains only the binary values. But, the input dataset (KDD cup-99) contains continuous variable for each attributes so that, the conventional algorithm is not suitable for mining frequent items. By considering this property, we simply find the 1-length items from each attributes by finding the frequency of the continuous variable present in each attribute and then, the frequent items are discovered by inputting the minimum support. These frequent items are identified for both class namely, normal and attack (combining four types of attacks).

(b) Identification of suitable attributes for rule generation

In this step, we have chosen only the most suitable attributes for identifying the classification whether the record is normal or attack. The reason behind this step is that the input data contain 34 attribute, in which all the attributes are not so effective in detecting the intrusion detection. For identifying the suitable attribute, we have used deviation method, where mined 1-length frequent items are used. At first, the mined 1-length items from each attribute are stored in a vector so that 34 vectors are obtained for each class (class 1 and class 2), represented as, $C_i = V_1, V_2, \dots, V_i, \dots, V_{34}$ where, $i = 1(refer \ to \ normal), 2(refer \ to \ attack)$. Each vector (V_i) contains frequent items, whose frequency is greater than $V_i = \{f_i ; 1 \le i \le m\};$ support. minimum $support(f_i) \ge min_supp$. Then, for each attribute, deviation range of frequent items is identified by comparing the frequent

items present within a vector such a way, the deviation range {max, min} is obtained for every vector.

$$D_{v(j)} = \{f_{\text{max}}, f_{\text{min}}\}; \text{ where, } f_{\text{max}} = Max(f_j); \quad f_{\text{min}} = Min(f_j)$$

Then, one-to-one comparison is performed in between both class of respective vector to identify the effective attribute. The attributes that not contain identical {max, min} range for both class is chosen as effective attribute, which will give significant detection rate rather than utilizing the all attribute for identifying the classification. The effective attributes chosen for rule generation process is represented as, $C_i = \left[V^{(1)}, V^{(2)}, \dots, V^{(j)}, \dots, V^{(k)}\right]$, Where, $k \le 34$.

(c) Rule generation

The effective attributes chosen from the previous step is utilized to generate rules that is derived from the {max, min} deviation. By comparing the deviation range of effective attributes in between the normal and attack data, the intersection points are identified for the effective attributes. By making use of these two intersection points, the definite and indefinite rules are generated. For example, {max, min} deviation range for normal data related to attribute1 is {1, 5} and {max, min} deviation for attack data corresponding to attribute1 is {2, 8}. Then, the rule is designed like, "IF attribute1 is greater than 5, THEN the data is attack, "IF attribute1 is in between 2 and 5, THEN the data is normal OR attack" and "IF attribute1 is less than 2, THEN the data is normal". In addition to that, some of the data contains only one intersection point, which provides only two rules.

(d) Rule filtering

In order to learn the fuzzy rules efficiently and design a compact and interpretable classification system, we should concentrate in these two criteria given in [33, 34]: (1) The number of fuzzy rules should be decreased as much as possible, (2) The IF part of fuzzy rules should be short. By concentrating on these two criteria, we have filtered the rules such a way that, we take only the short and less number of rules. The rules that are generated from the previous step contain definite and indefinite rules. The definite rules are the rules that contain only one classified label in the THEN part and indefinite rule contain two classification label data in the THEN part. The proposed rule filtering technique filters the indefinite rule and selects only the definite rules for learning the fuzzy system.

(e) Generating fuzzy rules

In general, fuzzy rules are defined within the fuzzy system manually or the rules are obtained from the domain expert. But, in the proposed system, we automatically find the fuzzy rules based on the mined 1-length frequent items. The fuzzy rules are generated from the definite rules, where the IF part of the rule is a numerical variable and THEN part is a class label related to attack name or normal. But, the fuzzy rule should contain only the linguistic variable. So, in order to make the fuzzy rules from the definite rules, we should fuzzify the numerical variable of the definite rules and THEN part of the fuzzy rule is same as the consequent part of the definite rules. For example, "IF attribute1 is H, THEN the data is attack and "IF attribute1 is VL, THEN the data is normal". These fuzzy rules are used to learn the fuzzy system so that the effectiveness of the proposed system will be improved rather than simply using the fuzzy rules without any proper techniques.

(3) Fuzzy Decision Module

This section describes the designing of fuzzy logic system for finding the suitable class label of the test dataset. Zadeh in the late 1960s [39] introduced Fuzzy logic and is known as the rediscovery of multivalued logic designed by Lukasiewicz. The designed fuzzy system shown in figure 2 contains 34 inputs and one output, where inputs are related to the 34 attributes and output is related to the class label (attack data or normal data). Here, thirty four-input, single-output of Mamdani fuzzy inference system with centroid of area defuzzification strategy was used for this purpose. Here, each input fuzzy set defined in the fuzzy system includes four membership functions (VL, L, M and H) and an output fuzzy set contains two membership functions (L and H). Each membership function used triangular function for fuzzification strategy. The fuzzy rules obtained from sub-section IV.B are fed to the fuzzy rule base for learning the system.

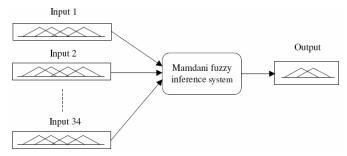


Figure 2. The designed Fuzzy system

(4) Finding an Appropriate Classification for a Test Input

For testing phase, a test data from the KDD-cup 99 dataset is given to the designed fuzzy logic system discussed in subsection IV.C for finding the fuzzy score. At first, the test input data containing 34 attributes is applied to fuzzifier, which converts 34 attributes (numerical variable) into linguistic variable using the triangular membership function. The output of the fuzzifier is fed to the inference engine which in turn compares that particular input with the rule base. Rule base is a knowledge base which contains a set of rules obtained from the definite rules. The output of inference engine is one of the linguistic values from the following set {Low and High} and then, it is converted by the defuzzifier as crisp values. The crisp value obtained from the fuzzy inference engine is varied in between 0 to 2, where '0' denotes that the data is completely normal and '1' specifies the completely attacked data.

V. EXPERIMENTATION

This section describes the experimental results and performance evaluation of the proposed system. The proposed system is implemented in MATLAB (7.8) and the performance of the system is evaluated using Precision, recall and F-measure. For experimental evaluation, we have taken KDD cup 99 dataset [37], which is mostly used for evaluating the performance of the intrusion detection system. For evaluating the performance, it is very difficult to execute the proposed system on the KDD cup 99 dataset since it is a large scale. Here, we have used subset of 10% of KDD Cup 99 dataset for training and testing. The number of records taken for testing and training phase is given in table III and table IV.

TABLE III. TRAINING DATASET TAKEN FOR EXPERIMENTATION

Training Dataset		
Normal	25,000	
DOS	25,000	
Probe	4107	
RLA	77	
URA	42	

TABLE IV. TRAINING DATASET TAKEN FOR EVALUATION

Testing Dataset		
Normal	26,000	
DOS	26,000	
Probe	4107	
RLA	77	
URA	42	

A. Experimental Results and Performance Analysis

The training dataset contains normal data as well as four types of attacks, which are given to the proposed system for identifying the suitable attributes. The selected attribute for rule generation process is given in table V. Then, using the fuzzy rule learning strategy, the system generates definite and indefinite rules and finally, fuzzy rules are generated from the definite rules.

TABLE V. SELECTED ATTRIBUTES FOR RULE GENERATION

Attribute Index	Selected Attributes
1	duration
5	src_bytes
6	dst_bytes
8	wrong_fragment
9	urgent
10	hot
11	num_failed_logins
13	num_compromised
16	num_root
17	num_file_creations
18	num_shells
19	num_access_files
23	count
24	srv_count

In the testing phase, the testing dataset is given to the proposed system, which classifies the input as a normal or attack. The obtained result is then used to compute overall accuracy of the proposed system. The overall accuracy of the proposed system is computed based on the definitions, namely precision, recall and F-measure which are normally used to estimate the rare class prediction. It is advantageous to accomplish a high recall devoid of loss of precision. F-measure is a weighted harmonic mean which evaluates the trade-off between them.

$$\begin{aligned} & Precision = \frac{TP}{TP + FP} \\ & Recall = \frac{TP}{TP + FN} \\ & F - measure = \frac{(\beta^2 + 1)(Precision \cdot Recall)}{\beta^2 \cdot Precision + Recall} \end{aligned} \text{ where, } \beta = 1 \\ & Overall\ accuracy = \frac{TP + TN}{TP + TN + FN + FP} \end{aligned}$$

Where, $TP \rightarrow$ True positive $TN \rightarrow$ True negative

$FN \rightarrow$ False negative

$FP \rightarrow$ False positive

These are computed using the confusion matrix in Table VI, and defined as follows:

TABLE VI. CONFUSION MATRIX

		Predicted class	
		Positive class	Negative class
Actual	Positive class	True positive (TP)	False negative (FN)
Class	Negative class	False positive (FP)	True negative (TN)

The evaluation metrics are computed for both training and testing dataset in the testing phase and the obtained result for all attacks and normal data are given in table VII, which is the overall classification performance of the proposed system on KDD cup 99 dataset. By analyzing the result, the overall performance of the proposed system is improved significantly and it achieves more than 90% accuracy for all types of attacks.

TABLE VII. THE CLASSIFICATION PERFORMANCE OF THE PROPOSED INTRUSION DETECTION SYSTEM

·	Metric	Proposed System		
	Metric	Training	Testing	
	Precision	0.912522	0.912522	
	Recall	0.37083	0.37083	
PROBE	F-measure	0.52735457	0.52735457	
	Accuracy	0.906208	0.909323	
	Precision	0.993563	0.993828	
	Recall	0.90144	0.904154	
DOS	F-measure	0.94526236	0.94687236	
	Accuracy	0.9478	0.949269	
	Precision	0.051948	0.051948	
	Recall	0.190476	0.190476	
U2R	F-measure	0.08163265	0.08163265	
	Accuracy	0.992812	0.993088	
	Precision	0.075949	0.075949	
	Recall	0.155844	0.155844	
R2L	F-measure	0.10212766	0.10212766	
	Accuracy	0.991586	0.991909	
	Precision	0.828439	0.829318	
NODMAT	Recall	0.99416	0.994385	
NORMAL	F-measure	0.90376539	0.90438129	
	Accuracy	0.910852	0.903019	

VI. CONCLUSION

We have developed an anomaly based intrusion detection system in detecting the intrusion behavior within a network. A fuzzy decision-making module was designed to build the system more accurate for attack detection, using the fuzzy inference approach. An effective set of fuzzy rules for inference approach were identified automatically by making use of the fuzzy rule learning strategy, which are more effective for detecting intrusion in a computer network. At first, the definite rules were generated by mining the single length frequent items from attack data as well as normal data. Then, fuzzy rules were identified by fuzzifying the definite rules and these rules were given to fuzzy system, which classify the test data. We have used KDD cup 99 dataset for evaluating the performance of the proposed system and experimentation results showed that the proposed method is effective in detecting various intrusions in computer networks.

REFERENCES

- [1] Yao, J. T., S.L. Zhao, and L.V. Saxton, "A Study On Fuzzy Intrusion Detection", In Proceedings of the Data Mining, Intrusion Detection, Information Assurance, And Data Networks Security, SPIE, Vol. 5812, pp. 23-30, Orlando, Florida, USA, 2005.
- [2] Nivedita Naidu and Dr.R.V.Dharaskar, "An Effective Approach to Network Intrusion Detection System using Genetic Algorithm", International Journal of Computer Applications, Vol.1, No.3, pp.26–32, February 2010.
- [3] J. Allen, A. Christie, and W. Fithen, "State Of the Practice of Intrusion Detection Technologies", Technical Report, CMU/SEI-99-TR-028, 2000
- [4] B.V. Dasarathy, "Intrusion Detection", Information Fusion, Vol.4, No.4, pp.243-245, 2003.
- [5] R.G.Bace, "Intrusion Detection", Macmillan Technical Publishing, Indianapolis, USA, 2000.
- [6] Marcos M. Campos, Boriana L. Milenova, "Creation and Deployment of Data Mining-Based Intrusion Detection Systems in Oracle Database 10g", in Proceedings of the Fourth International Conference on Machine Learning and Applications, 2005.
- [7] Anazida Zainal, Mohd Aizaini Maarof and Siti Maryam Shamsudin, "Research Issues in Adaptive Intrusion Detection", in Proceedings of the 2nd Postgraduate Annual Research Seminar (PARS'06), Faculty of Computer Science & Information Systems, Universiti Teknologi Malaysia, 24 – 25 May, 2006.
- [8] Dr. Fengmin Gong, "Deciphering Detection Techniques: Part II Anomaly-Based Intrusion Detection", White Paper from McAfee Network Security Technologies Group, 2003.
- [9] Susan M. Bridges and Rayford B.Vaughn, "Fuzzy Data Mining And Genetic Algorithms Applied To Intrusion Detection", In Proceedings of the National Information Systems Security Conference (NISSC), Baltimore, MD, pp.16-19, October 2000.
- [10] Jian Pei, Upadhyaya, S.J., Farooq, F., Govindaraju, V, "Data mining for intrusion detection: techniques, applications and systems ", in Proceedings of the 20th International Conference on Data Engineering, pp: 877 - 87, 2004.
- [11] Cannady J, "Artificial Neural Networks for Misuse Detection", in Proceedings of the '98 National Information System Security Conference (NISSC'98), pp. 443-456, 1998.
- [12] Shon T, Seo J, and Moon J, "SVM Approach with A Genetic Algorithm for Network Intrusion Detection", Lecture Notes in Computer Science, Springer Berlin / Heidelberg, Vol. 3733, pp. 224-233, 2005.
- [13] Yu Y, and Huang Hao, "An Ensemble Approach to Intrusion Detection Based on Improved Multi-Objective Genetic Algorithm", Journal of Software, Vol.18, No.6, pp.1369-1378, June 2007.

- [14] J. Luo, and S. M. Bridges, "Mining fuzzy association rules and fuzzy frequency episodes for intrusion detection", International Journal of Intelligent Systems, Vol. 15, No. 8, pp. 687-704, 2000.
- [15] W. Lee, S. Stolfo, and K. Mok, "A Data Mining Framework for Building Intrusion Detection Model", In Proceedings of the IEEE Symposium on Security and Privacy, Oakland, CA, pp. 120-132, 1999.
- [16] Dewan Md. Farid and Mohammad Zahidur Rahman, "Anomaly Network Intrusion Detection Based on Improved Self Adaptive Bayesian Algorithm", Journal of Computers, Vol.5, No.1, January, 2010.
- [17] K.Yoshida, "Entropy Based Intrusion Detection", in Proceedings of the IEEE Pacific Rim Conference on Communications, Computers and signal Processing, Vol. 2, pp. 840 – 843, Aug 28-30, 2003.
- [18] Sujaa Rani Mohan, E.K. Park, Yijie Han, "An Adaptive Intrusion Detection System Using A Data Mining Approach", White paper from University of Missouri, Kansas City, October 2005.
- [19] Rasha G. Mohammed Helali, "Data Mining Based Network Intrusion Detection System: A Survey", In Novel Algorithms and Techniques in Telecommunications and Networking, pp. 501-505, 2010.
- [20] Pakkurthi Srinivasu, P.S. Avadhani, Vishal Korimilli, Prudhvi Ravipati, "Approaches and Data Processing Techniques for Intrusion Detection Systems", Vol. 9, No. 12, pp. 181-186, 2009.
- [21] G. Macia Fernandez and E. Vazquez, "Anomaly-based network intrusion detection: Techniques, systems and challenges", Computers & Security, Vol. 28, No. 1-2, pp. 18-28, February-March 2009.
- [22] Mark Crosbie and Gene Spa Ord, "Defending a Computer System using Autonomous Agents", Technical report, 1995.
- [23] Honig, A., Howard, A., Eskin, E., and Stolfo, S. J., "Adaptive Model Generation: An Architecture for the Deployment of Data Mining-Based Intrusion Detection Systems, Applications of Data Mining in Computer Security, Kluwer Academic Publishers, Boston, MA, pp. 154-191, 2002.
- [24] Stephen F. Owens, Reuven R. Levary, "An adaptive expert system approach for intrusion detection", International Journal of Security and Networks, Vol. 1, No. 3/4, pp. 206-217, 2006.
- [25] Alok Sharma, Arun K. Pujari, Kuldip K. Paliwal, "Intrusion detection using text processing techniques with a kernel based similarity measure", Computers & Security, Vol. 26, No. 7-8, pp. 488-495, 2007.
- [26] Shi-Jinn Horng, Ming-Yang Su, Yuan-Hsin Chen, Tzong-Wann Kao, Rong-Jian Chen, Jui-Lin Lai, Citra Dwi Perkasa, "A novel intrusion detection system based on hierarchical clustering and support vector machines", Expert Systems with Applications, Vol. 38, No. 1, pp. 306-313, 2011.
- [27] Abadeh, M.S., Habibi, J., "Computer Intrusion Detection Using an Iterative Fuzzy Rule Learning Approach", in Proceedings of the IEEE International Conference on Fuzzy Systems, pp: 1-6, London, 2007.
- [28] Bharanidharan Shanmugam, Norbik Bashah Idris, "Improved Intrusion Detection System Using Fuzzy Logic for Detecting Anamoly and Misuse Type of Attacks", in Proceedings of the International Conference of Soft Computing and Pattern Recognition, pp: 212-217, 2009.
- [29] O. Adetunmbi Adebayo, Zhiwei Shi, Zhongzhi Shi, Olumide S. Adewale, "Network Anomalous Intrusion Detection using Fuzzy-Bayes", IFIP International Federation for Information Processing, Vol. 228, pp. 525-530, 2007.
- [30] Arman Tajbakhsh, Mohammad Rahmati, Abdolreza Mirzaei, "Intrusion detection using fuzzy association rules", Applied Soft Computing, Vol. 9, No. 2, pp. 462-469, 2009.
- [31] Zhenwei Yu, Tsai, J.J.P., Weigert, T., "An Automatically Tuning Intrusion Detection System", IEEE Transactions on Systems, Man, and Cybernetics, Vol. 37, No. 2, pp. 373 - 384, 2007.
- [32] Qiang Wang and Vasileios Megalooikonomou, "A clustering algorithm for intrusion detection", in Proceedings of the conference on Data Mining, Intrusion Detection, Information Assurance, and Data Networks Security, vol. 5812, pp. 31-38, March 2005.
- [33] Cordon O, Gomide F, Herrera F, Hoffmann F, Magdalena L, "Ten years of genetic fuzzy systems: current framework and new trends", Fuzzy Sets and Systems, vol.141, no.1, pp. 5–31, 2004.
- [34] M. Saniee Abadeh, J. Habib and C. Lucas, "Intrusion detection using a fuzzy genetics-based learning algorithm", Journal of Network and Computer Applications, vol.30, no.1, pp. 414–428, 2007.

- [35] R. Agrawal, T. Imielinski, A., Swami, "Mining association rules between sets of items in large databases", in Proceedings of 1993 ACM SIGMOD Intl. Conf. on Management of Data, Washington, DC, pp. 207–216, 1993.
- [36] http://www.ll.mit.edu/mission/communications/ist/corpora/ideval/data/ 1998data html
- $[37] \ http://www.sigkdd.org/kddcup/index.php?section=1999\&method=data$
- [38] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu and Ali A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set", in Proceedings of the Second IEEE international conference on Computational intelligence for security and defense applications, pp. 53-58, Ottawa, Ontario, Canada, 2009.
- [39] Zadeh, L.A., "Fuzzy sets", Information and control, vol.8, pp. 338-353, 1965
- [40] Jiawei Han, Jian Pei, Yiwen Yin, Runying Mao, "Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach", Data Mining and Knowledge Discovery, Vol. 8, No. 1, pp. 53 - 87, 2004.



R. Shanmugavadivu received her B.Sc. and M.Sc (computer science) degrees from the Department of Computer Science, PSG College of Arts & Science, Coimbatore Affiliated to Bharathiyar University in 1995 and 1998, respectively. Completed her M.PHIL degree in Computer science in 2008, at Bharathiyar University, Coimbatore .She is currently working as Assistant professor, Department of Computer Science at PSG

College of Arts & Science, Coimbatore.



Dr.N.Nagarajan, received his M.E., and B.Tech., degrees in the disciplines of Electronics Engineering from Madras Institute of Technology, under Anna University, Chennai in 1984 and 1982 respectively. He Obtained Ph.D. degree from Anna University, Chennai in "Faculty of Information and Communication Engineering" in 2006..He posses 25 years of teaching experience in various reputed Engineering colleges viz., Kongu Engineering College, Sri Krishna College of

Engineering and Technology, Sri Ramakrishna Engineering College etc., .Currently, he is working as a Principal of Coimbatore Institute of Engineering and Information Technology, Coimbatore. He has Published 15 papers in the International refereed journals and 20 papers in International and National conferences. He is a Reviewer for WSEAS International Transactions. He had received a grant of Rupees Five Lakhs in MODROB scheme from AICTE, New Delhi for modernizing Communication Laboratory using Fibre Optic Communication at Kongu Engineering College, Perundurai, Erode, Tamilnadu during the year 1999.He was Selected in, "2000 Outstanding Scientists" for the year 2008-2009 by International Biographical Centre, Great Britain in its 34th edition. He was also nominated as "International Scientist of the year "for 2008 by International Biographical Centre, Cambridge, England.

Blemish Tolerance in Cellular Automata And Evaluation Reliability

Roghayyeh parikhani Engineering Department, Islamic Azad University, Tabriz branch Tabriz, Iran

Mohmad teshnelab

Department of Controls Engineering, Faculty of Electrical and Computer Engineering,KN Toosi University of Technology Tehran, Iran Shahram babaei Engineering Department, Islamic Azad University, Tabriz branch Tabriz, Iran

Abstract—The computational paradigm known as quantum-dot cellular automata (QCA) encodes binary information in the charge configuration of Coulomb-coupled quantum-dot cells. Functioning QCA devices made of metal-dot cells have been fabricated and measured. We focus here on the issue of robustness in the presence of disorder and thermal fluctuations. We examine the performance of a semi-infinite QCA shift register as a function of both clock period and temperature. The existence of power gain in QCA cells acts to restore signal levels even in situations where high speed operation and high temperature operation threaten signal stability. Random variations in capacitance values can also be tolerated.

Keywords-component; QCA, molecular electronics, single electronics, quantum-dot cellular automata, nanoelectronics

I. INTRODUCTION

Conventional transistor-based CMOS technology faces

great challenges with the down-scaling of device sizes in

recent years. Issues such as quantum effects, dopant-induced disorder, and power dissipation may hinder further progress in scaling microelectronics. As the scaling approaches a molecular level, a new paradigm beyond using current switches to encode binary information may be needed. Quantum-dot cellular automata (QCA) [1–3, 6, 12, 13, 18, 21] emerges as one such a paradigm. In the QCA approach bit information is

encoded in the charge configuration within a cell. Columbic interaction between cells is sufficient to accomplish the computation; no current flows out of the cell. It has been shown that very low power dissipation is possible [8].

A clocked QCA cell constructed with six quantum dots is shown in Fig. 1. Dots are simply places where a charge is localized. Two mobile electrons are present in the cell. The electrons will occupy antipodal sites in the corner dots because of Coulomb repulsion. The two configuration states correspond to binary information of "1" and "0" The electrons can also be pulled to middle dots if the occupancy energy in the middle dots is lower than corner dots. In this case we term the configuration "null" with no binary information present. The clock adjusts the relative occupancy energy between active dots

in the corner and null dots in the middle, pushing electrons to either active dots or null dots. The cell therefore switches between null state and active state. When a cell is placed close to another cell (as shown in Fig. 1b), they will have the same polarization due to Coulomb coupling. Based on the cell-to-cell interaction, logical QCA devices like binary wires, inverters, majority gates and full adders can all be implemented [18].

QCA devices exist. QCA devices made of metal-dot cells have been successfully demonstrated at low temperatures. Majority gates, binary wires, memories, clocked shift registers and fan outs have all been fabricated [1–3, 12, 13, 21]. Figure 2 shows a schematic diagram and scanning electron micrograph of a clocked shift register. Aluminum islands form the dots and Al/AlOx tunnel junctions serve as the tunneling path between dots.

Tunnel junctions are fabricated with shadow evaporation technique. Multiple tunnel junctions are used instead of a single junction to suppress co-tunneling. The clock is implemented by simply applying voltage to leads capacitively coupled to the middle dots. Single electron transistors (SET_s) are used as readout electrometers. Though the operation of metal-dot QCA devices is restricted to cryogenic temperatures, they may be viewed as prototypes for molecular QCA cells that will operate at room temperature. It may well be that molecular QCA,with the possibility of enormous functional densities, very low power dissipation, and room temperature operation, is finally the most promising system [5, 9–11, 14, 16].

Metal-dot QCA do have the advantage of having been already created and tested, and we expect that understanding the details of robustness in the metal-dot system will yield benefits for designing molecular systems. Here we focus on the robustness in metal-dot QCA circuits. In particular, we consider theoretically the effects of temperature, random variations in capacitance, and operating speed, on the performance of a semiinfinite QCA shift register. The paper is organized as follows: in Section II, we describe the application of single-electron tunneling theory to metal QCA devices.

Section III describes the characterization of power gain in QCA circuits. In Section IV we analyze the operation of a semi-infinite QCA shift register. Finally, in Section V we

and non-leaky capacitors.

a active

null

"1"

b

"0"

Fig. 1. Schematic of a QCA cell. a The three states of a single cell. b Coulomb interactions couple the states of neighboring cells.

lculate behavior of the QCA shift register in the limits of high speed, high temperature, and high defect levels.

II. SINGLE ELECTRON SYSTEM THEORY

Metal-dot QCA can be described with the orthodox theory of coulomb blockade [19]. The circuit is defined by charge configurations, which are determined by the number of electrons on each of the metal islands. Metal islands are regions of metal surrounded by insulators; at zero temperature they hold an integer number of charges. The islands play the role of QCA dots and are coupled to other islands and leads through tunnel junctions (i.e., quantummechanically leaky capacitors)

Leads by contrast are metal electrodes whose voltages are fixed by external sources. We define dot charge qi as the charge on island i and qk 0 as the charge on lead k. The free energy of charge configuration within the circuit is the electrostatic energy stored in the capacitors and tunnel junctions minus the work done by the voltage sources [20]:

$$F = \frac{1}{2} \begin{bmatrix} q \\ q' \end{bmatrix}^T C^{-1} \begin{bmatrix} q \\ q' \end{bmatrix} - v^T q' \tag{1}$$

Here C is the capacitance matrix including all the junctions and capacitors, v is the column vector of lead voltages, and q and q0 are the column vectors of dot charges and lead charges. At zero temperature, the equilibrium charge configuration is the one that has minimum free energy and the number of charges on each islands is exactly an integer. A tunneling event happens at zero temperature only if the free energy is lower for the final state than for the initial state. At finite temperatures, a dot charge need no longer be an integer but is rather a thermal average over all possible configurations. A thermally excited tunneling event may happen even when the

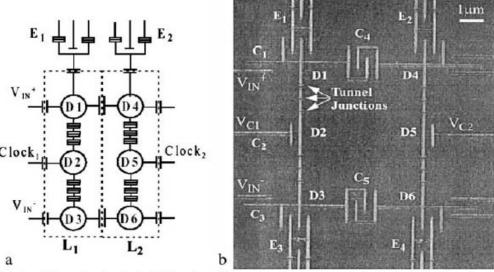
free energy increases. The transition rate of tunneling between two charge configuration states at a certain temperature T is given by

$$\Gamma_{ij} = \frac{1}{e^2 R_{\rm T}} \frac{\Delta F_{ij}}{1 - e^{-\Delta F_{ij}/(kT)}}$$
 (2)

where RT is the tunneling resistance, ΔF_{ij} is the energy difference between the initial state i and final state j.

The tunneling events can be described by a master equation—a conservation law for the temporal change of the probability distribution function of a physical quantity,

$$\frac{\mathrm{d}\boldsymbol{P}}{\mathrm{d}t} = \Gamma \boldsymbol{P} \tag{3}$$



ca

Fig. 2. a Schematic of a clocked shift register. b Scanning electron micrograph of a clocked shift register.

where P is the vector of state probabilities and Γ is the transition matrix. From the solution P(t) we can obtain the ensemble average of the charge on each dot. We solve Eq. 3 directly and find the dot charge as a function of time; from this we can obtain any other voltage or charge in the circuit. In many systems direct solution of the master equation, which requires the enumeration of all the accessible states of the system is impractical due to the large set of accessible states. Because QCA operates so near the instantaneous ground state of the system, complete enumeration of the accessible states is possible and we need not resort to Monte Carlo methods.

III. POWER GAIN IN QCA

A robust circuit must have power gain in order to restore signals weakened due to unavoidable dissipative processes. In conventional CMOS, the power supply provides the energy power gain. In QCA systems the energy needed for power gain is supplied by the clock. A weak input is augmented by work done by the clock to restore logic levels. Power gain has been studied theoretically in molecular QCA circuits [8] and measured experimentally in metal-dot QCA circuits [3]. Power gain is defined by the ratio of the work done by the cell on its neighbor to the right (the output of the cell), to the work done on the cell by its neighbor to the left (the input to the cell). The work done on a cell by an input lead coupled through an input capacitor C over a time interval T is given by

$$W = \int_{0}^{T} V(t) \frac{\mathrm{d}}{\mathrm{d}t} Q_{c}(t) \mathrm{d}t$$
 (4)

where V(t) is the lead voltage, Qc(t) is the charge on the input capacitor. We consider the total work done over a clock period so the cell configuration is the same at t=0 and t=T. The power gain is thus the ratio of output to input signal power W_{out}/W_{in} , where each sums the work done by (on) all input (output) leads.

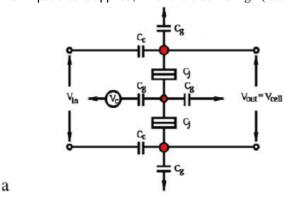
IV. OPERATION OF SEMI-INFINITE QCA SHIFT REGISTER

The schematic of a clocked half QCA cell is shown in Fig.3a. The capacitances are taken to be Cj=1.6 aF, Cg=0.32 aF, Cc=0.8 aF, and the tunneling resistance RT=100 kW. Eachisland is grounded through a capacitance of 0.32 aF. These are physically reasonable though somewhat better (meaning capacitances are smaller) than the experiments have so-far achieved. Input is applied to the top and bottom dot through coupling capacitors. The potential difference between the top and bottom dots is the output Vcell.

The phase diagram of the equilibrium charge state configuration of the cell shown in Fig. 3a is plotted in Fig. 4. The diagram shows the calculated stable regions of charge configuration as a function of input and clock potential. Each hexagonal region is labeled by three integers (n1, n2, n3), the number of elementary charges in the top, middle, and bottom dot, respectively. A positive number indicates an extra hole and negative number represents an extra electron. Each hexagon represents the configuration state that has, for those

values of input voltage and clock voltage, the lowest free energy.

The clocking cycle can be envisioned as follows. First, a small input bias is applied, when the clock is high (less



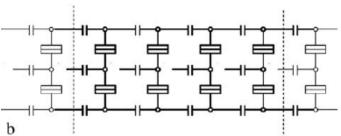


Fig. 3. a Schematic of a clocked triple dot. The input is applied to the top and bottom dot. The clock is set to the middle dot. The output defined as Vcell is the differential potential etween the top and the bottom dot. Cj=1.6 aF, Cg=0.32 aF, Cc=0.8 aF. The capacitor to ground is 0.32 aF. RT=100 kW. b Schematic of a shift register composed of a line of identical triple dots in a. The thick line described the actual four cells simulated.

negative, in fact for this circuit 0). This situation corresponds to point a in Fig. 4; no electron switching event happens and the cell remains in the null state, holding no information. When the clock is then lowered (more negative) the system moves along the line shown through point b. An electron is switched to either top dot or bottom dot, decided by the input; the cell is then in the active state. If the clock is held very negative (point c), the electron is locked in the active state, since the energy barrier in the middle dot is too high to overcome. The locked cell is essentially a single bit memory—its present state depends on its state in the recent past, not on the state of neighbors. Varying clock potential gradually between point a and c will switch the cell between null, active and locked state adiabatically.

A QCA shift register can be constructed with a line of capacitively coupled half QCA cells shown in Fig. 3b, where the output from each cell acts as the input to its right neighbor. The transport of information from cell to cell is controlled by clock signals. Initially, all the cells are in the null state since the clocks are high. Then an input signal is applied to the first cell and the clock for the first cell is lowered. The first cell thus switches to the opposite state of the input and holds to that state even when input is removed.

When the clock for the second cell is lowered, the second

cell switches to the opposite state to the first cell accordingly and locks the bit. The information is thereafter propagated along the cell line by the clock signals. Each cell in turn copies (an inverts) a bit from its neighbor to the left when the left neighbor is in the locked state and erases the bit, i.e., returns to the null state, while its right neighbor still holds a copy (inverted) of the bit. The copying of the bit can be accomplished gradually so that the switching cell is always close to its instantaneously ground state and thus dissipates very little energy.

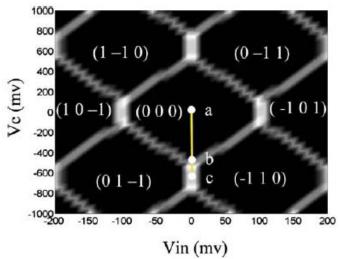


Fig. 4. The equilibrium state configuration of a triple dot cell described in Fig. 6. (n1, n2, n3) are the number of charges in the top, middle and bottom dot, respectively. The cell is in the null state in point a. The cell is in the active state in point b. The cell is in locked state in point c.

It_s instructive to model a semi-infinite shift register in order to study the robustness in the QCA circuit. A four phase clocking scheme is adopted to achieve adiabatic switching, shown in Fig. 5. Each clock signal is shifted a quarter-period. As a bit moves down the shift register, we need model only a four QCA half-cells at a time, since by the time the bit is latched in the leading cell, the leftmost cell has returned to null. This is equivalent to viewing the simulation as occurring on a ring of four half-cells. Figure 6 shows the time evolution of cell potentials for four neighboring cells in a semi-infinite shift register. The shaded areas indicate stored bit information. Each cell has the opposite signal to its neighboring cells with a quarter period shift; the information is both copied and inverted.

The arrow indicates the direction of the information flow. At the end of the first quarter clock period, the first clock is set to lowso that the first cell latches the input and locks it while the second cell is in the null state. By the time the second clock is low, the first cell is still kept locked. The second cell thus copies the bit from the first cell. By the end of the third quarter period, the bit in the first cell is erased as its clock is set to high. The third cell copies the bit from the second cell and holds it. The process goes on and the bit information is transported along the chain. Note that there are always at least two copies of the bit at one time. When there are three copies

of the bit, the cell potential in the middle cell decreases slightly (in absolute value) while the cell potential in its left and right neighbor increase slightly (thus the small Bnotch^ in the center of the flat parts of the waveform).

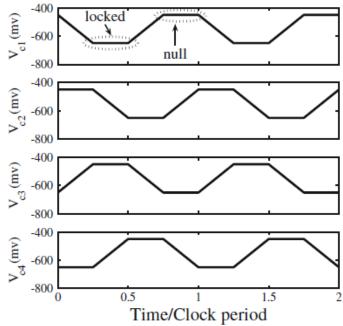


Fig. 5. A four phase clocking scheme in metal-dot QCA.

V. OPERATION OF SEMI-INFINITE QCA SHIFT REGISTER

A. EFFECT OF TEMPERATURE AND SPEED

Because of the difficulty of fabricating small capacitors, metal-dot QCA circuits operate at low temperatures.

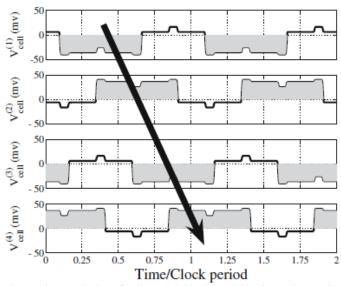


Fig. 6. Time evolution of cell potential in the neighboring cells. Vcell (n) is the differential potential between the top and the bottom dot of the nth cell.

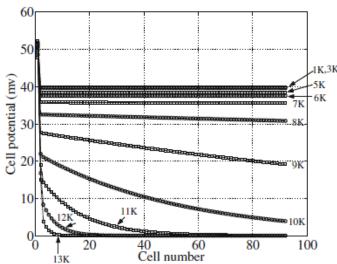


Fig. 8. Deviation from unity power gain for an individual cell as a function of temperature

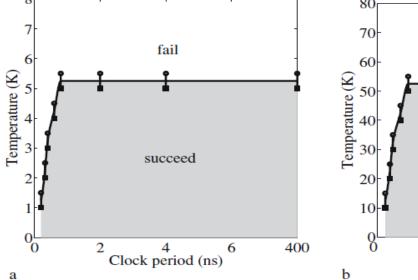
Thermal excitation is therefore a potential source of random error in metal-dot QCA circuits, and it is clear that at high enough temperatures the circuit will fail. It is tempting to conclude that for a long line of cells, failures are unavoidable at any non-zero temperature. It is well known that there is no long-range order in one-dimensional systems [15].

While the energy for a mistake might be higher than kBT, the degeneracy (and therefore entropy) of mistake states increases as the system size expands. For a system in thermal equilibrium therefore, the free energy of the mistake states eventually become lower than the mistake-free zero-entropy ground state [7]. A static (unclocked) chain of QCA cells therefore has, for any non-zero temperature, a characteristic length (_ eEk=kBT) after which mistakes become very likely. But a clocked line is not in thermal equilibrium—it is actively driven. The clock can supply energy to the system to restore

signal states.

To see the effect of temperature on the performance of the clocked semi-infinite shift register, we here solve the timedependent problem of the clocked shift register using the master-equation (Eq. 3) approach described in Section II. The calculated cell potential (see Fig. 3) of the kth cell in the chain at time t is Vcell(k,t). When each cell in the chain in turn latches the bit the cell potential is at its largest magnitude. Figure shows this maximum cell potential Vcell(k)=max(|Vcell(k,t)|) as a function of cell number k down the chain. The calculated response is plotted for various values of the temperature. The cell potential is higher at the very beginning of the chain simply because the first cell is driven by an input voltage which is a stronger driver than subsequent cells see; they are driven by other cells. At temperatures above 10 K the cell potential decays with distance as information is transported along the chain. At each stage the signal deteriorates further, and for a long shift register the information will be lost. For individual cells, this means errors due to thermal fluctuations become increasingly more likely. As the temperature is lowered the signal decay-length increases. At temperatures below 5 K, however, the behavior appears qualitatively different— the cell potential remains constant along the long the chain.

To the accuracy of our calculation for a large but finite number of cells, no signal degradation appears at all. The degradation of performance with increasing temperature can be explained in terms of power gain. We calculate the power gain of each individual cell in the chain by directly calculating the work done on the cell by its neighbor to the left, and the work done by the cell on its neighbor to the right. For each operating temperature the power gain is the same for each cell (apart from those very near the beginning of the line). If the power gain is precisely 1 (or greater), then there is no signal degradation moving down the line. At each cell, power is drawn from the clock sufficient to completely restore the signal as it is copied to the next cell. We refer to the situation in



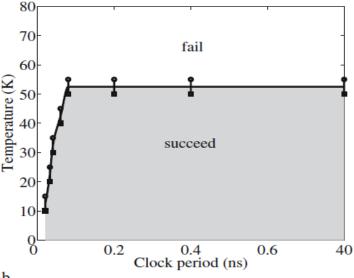
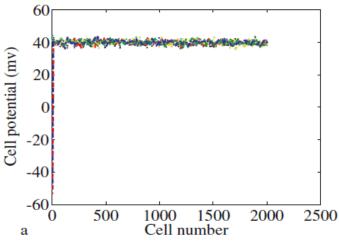


Fig. 9. The phase diagram of the operation space as a function of temperature and clock period when a. Cj=1.6 aF, and b. Cj=0.16 aF. The shaded area below the curve is where the circuit succeeds and the white area is where the circuit fails.



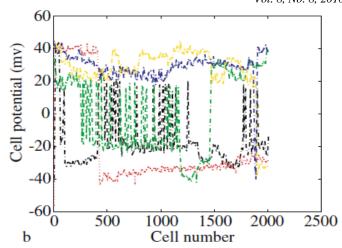


Fig. 10. Cell potential as a function of cell number at 4 K when capacitance variation is a ±10% b ±15%.

which unity power gain enables transmission of signals over arbitrarily long distances as Brobust[^] operation. If the power gain is less than 1, then the signal will be degraded as it moves down the line. Figure 8 shows the deviation from unity power gain as a function of temperature on a logarithmic scale. For temperatures below 5 K the power gain is 1; above 5 K, the power gain is less than 1. At higher temperatures, the flow of energy from the clock can no longer compensate for the energy loss to the thermal environment, with the result that the signal decays at each stage. As the difference between the power gain and 1 becomes small our analysis is limited by the numerical accuracy of the calculation. Nevertheless, the exponential character of the approach to unity power gain supports the interpretation that this transition is a qualitative change between robust and non-robust behavior, analogous to a phase transition.

The time-dependent calculation above is repeated for various temperatures and clock speeds to generate the phase diagram of the operational space of the circuit shown in Fig. 9. We display the results for the circuit with our standard parameters, with Cj=1.6 aF in Fig. 9a and for more aggressively scaled parameters, with Cj=0.16 aF in Fig. 9b. All capacitances and voltages in the circuit are scaled appropriately with Cj. The aggressively scaled parameter calculation illustrates scalability of QCA circuits.

The performance of the circuit will increase with smaller capacitances. The shaded area below the curve indicates speeds and temperatures for which the circuit is robust. The white area represents non-robust operation for which bit information decays along the chain. The two figures are identical except for the scale: the aggressively scaled circuit of Fig. 9b operates ten times faster and at a temperature ten times higher than the circuit in Fig. 9a. The area of robust operation is limited by both speed and temperature. In Fig. 9a, when the clock period is less than about 0.2 ns (corresponding to 5 GHz), the circuit fails (is not robust) even at zero temperature. This occurs as the clock period approaches the electron tunneling rate. When the clock speed is too fast, the electrons do not have enough time to tunnel reliably from one dot to another. The error probability accumulates as the information moves along the chain.

Increasing the clock period increases the probability of electrons being in the Bright^ states. This improvement quickly saturates and further increasing the clock period has no effect since the electrons have had enough time to be in the correct state.

The tunneling rate is related to the tunnel resistance, so this description is equivalent to the observation that the speed is limited to the RC time-constant of the circuit.

B. DEFECT TOLERANCE IN THE QCA SHIFT REGISTER

A robust circuit must be tolerant of defects that introduce variations in the values of the designed parameters. We consider the situation of a very long shift register in which the value of each capacitor in the circuit is varied randomly within a fixed percentage range from its its nominal value. The circuit is robust if the perturbation of the capacitances does not influence the performance of the circuit. We choose a working point in Fig. 9a where clock period is 5 ns, the temperature is 4 K, and vary all the capacitances randomly by T10 and T15%. Figure 10 shows the cell potential as a function of cell number with random capacitance variation. Different color represents different capacitance variation within the certain percentage range. When the deviation is T10%, the circuit is robust and transmits bit information with no errors. The bit information is carried on correctly even at the 2,000th cell. When the deviation increases to T15%, the circuit is fragile since cells are flipped to the wrong states during propagation. This calculation demonstrates, again as a result of the power gain in each cell, that QCA circuits can tolerate considerable variation in parameter values and still function correctly.

VI. CONCLUSION

The QCA approach represents an entirely new way of encoding, moving, and processing binary information. As more experimental realizations of devices appear, attention naturally turns to the broader circuit behavior of these new devices. While molecular QCA may represent the most realistic long-

term system for robust room temperature operation, the metaldot QCA system provides an extremely valuable prototype system in which to explore QCA properties. Metal dot systems also have the advantage of being realizable now.

We have explored here the behavior of metal-dot QCA systems under stress—stressed by high temperature operation, high speed operation, and random variation in parameter values. In each case enough stress destroys the correct operation of the circuit. What we observe however is that these systems are not terribly fragile, they can survive in a broad range of operational space. In each case small errors threaten to accumulate over many cells and result in signal loss. The key feature is power gain from the clocking circuit which provides considerable robustness against these error mechanisms, restoring signal levels at each stage

REFERENCES

- I. Amlani, A. Orlov, G. Toth, G.H. Bernstein, C.S. Lent, and G.L.Snider, Science, vol. 284, p. 289, 1999.
- [2] R.K. Kummamuru, J. Timler, G. Toth, C.S. Lent, R.Ramasubramaniam, A. O. Orlov, G.H. Bernstein, and G.L. Snider, Appl. Phys. Lett., vol. 81.
- [3] R.K. Kummamuru, A.O. Orlov, C.S. Lent, G.H. Bernstein, and G.L.Snider, IEEE Trans. Electron Devices, vol. 50, pp. 1906–1913, 2003.
- [4] R.K. Kummamuru, M. Liu, A.O. Orlov, C.S. Lent, G.H. Bernstein, and G.L. Snider, Microelectron. J., vol. 36, 2005.
- [5] C.S. Lent and B. Isaksen, IEEE Trans. Electron Devices, vol. 50,pp. 1890–1896, 2003.
- [6] C.S. Lent, P.D. Tougaw, W. Porod, and G.H. Bernstein, Nanotechnology, vol. 4, p. 49, 1993.

- [7] C.S. Lent, P.D. Tougaw, andW. Porod, PhysComp_94, The Proceedings of the Workshop on Physics and Computing, pp. 5– 13, Dallas, TX:IEEE Computer Society Press, Nov. 17–20 1994.
- [8] C.S. Lent, B. Isaksen, and M. Lieberman, J. Am. Chem. Soc., vol.125, pp. 1056–1063, 2003.
- [9] Z. Li and T.P. Fehlner, Inorg. Chem., vol. 42, pp. 5715–5721, 2003.
- [10] Z. Li, A.M. Beatty, and T.P. Fehlner, Inorg. Chem., vol. 42, pp. 5715–5721, 2003.
- [11] M. Lieberman, S. Chellamma, B. Varughese, Y.L. Wang, C.S. Lent, G.H. Bernstein, G.L. Snider, and F.C. Peiris, Ann. N.Y. Acad. Sci., vol. 960, pp. 225–239, 2002.
- [12] A.O. Orlov, I. Amlani, G.H. Bernstein, C.S. Lent, and G.L. Snider, Science, vol. 277, p. 928, 1997.
- [13] A.O. Orlov, I. Amlani, R.K. Kummamuru, R. Ramasurbramaniam,G. Toth, C.S. Lent, G.H. Bernstein, and G.L. Snider, Appl. Phys.Lett., vol. 77, pp. 295–297, 2000.
- [14] H. Qi, S. Sharma, Z. Li, G.L. Snider, A.O. Orlov, C.S. Lent, and T.P. Fehlner, J. Am. Chem. Soc., vol. 125, pp. 15250–15259, 2003.
- [15] D.J. Thouless, Phys. Rev., vol. 187, pp. 732-733, 1969.
- [16] J. Timler and C.S. Lent, J. Appl. Phys., vol. 91, pp. 823–832, 2002.
- [17] G. Toth and C.S. Lent, J. Appl. Phys., vol. 85, pp. 2977–2984, 1999.
- [18] P.D. Tougaw and C.S. Lent, J. Appl. Phys., vol. 75, no. 3, pp. 1818– 1825, 1994.
- [19] C. Wasshuber, Computational single-electronics, Berlin HeidelbergNew York: Springer, 2001.
- [20] C. Wasshuber, H. Kosina, and S. Selberherr, IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst., vol. 16, p. 9, 1997.
- [21] K.K. Yadavalli, A.O. Orlov, R.K. Kummamuru, C.S. Lent, G.H.Bernstein, and G.L. Snider, BFanout in Quantum-dot CellularAutomata, 63rd Device Research Conference, Santa Barbara, CA, 2005.

Feed Forward Neural Network Algorithm for Frequent Patterns Mining

Amit Bhagat¹
Department of Computer Applications

Dr. Sanjay Sharma²
Associate Prof. Deptt. of Computer Applications

Dr. K.R.Pardasani³
Professor Deptt. of Mathematics

Maulana Azad National Institute of Technology, Bhopal (M.P.)462051, India

Abstract: Association rule mining is used to find relationships among items in large data sets. Frequent patterns mining is an important aspect in association rule mining. In this paper, an efficient algorithm named Apriori-Feed Forward(AFF) based on Apriori algorithm and the Feed Forward Neural Network is presented to mine frequent patterns. Apriori algorithm scans database many times to generate frequent itemsets whereas Apriori-Feed Forward(AFF) algorithm scans database Only Once. Computational results show the Apriori-Feed Forward(AFF) algorithm performs much faster than Apriori algorithm.

Keywords: Association rule mining, dataset scan, frequent itemsets, Neural Network..

I. INTRODUCTION

Data mining has recently attracted considerable attention from database practitioners and researchers because it has been applied to many fields such as market strategy, financial forecasts and decision support [1]. Many algorithms have been proposed to obtain useful and invaluable information from huge databases [2]. One of the most important algorithms is mining association rules, which was first introduced in [3, 4]. Association rule mining has many important applications in our life. An association rule is of the form $X \Rightarrow Y$. And each rule has two measurements: support and confidence. The association rule mining problem is to find rules that satisfy user-specified minimum support and minimum confidence. It mainly includes two steps: first, find all frequent patterns; second, generate association rules through frequent patterns. Many algorithms for mining association rules from transactions database have been proposed [5, 6, 7]since Apriori algorithm was first presented. However, most algorithms were based on Apriori algorithm which generated and tested candidate itemsets iteratively. This may scan

database many times, so the computational cost is high. In order to overcome the disadvantages of Apriori algorithm and efficiently mine association rules without generating candidate itemsets, many authors developed some improved algorithms and obtained some promising results [9,10, 11, 12, 13]. Recently, there are some growing interests in developing techniques for mining association patterns without a support constraint or with variable supports [14, 15, 16]. Association rule mining among rare items is also discussed in [17,18]. So far, there are very few papers that discuss how to combine Apriori algorithm and Neural Network to mine association rules. In this paper, an efficient algorithm named Apriori-Feed Forward(AFF) based on Apriori algorithm and Feed Forward Neural Network is proposed, this algorithm can efficiently combine the advantages of Apriori algorithm and Structure of Neural Network. Computational results verify the good performance of the Apriori-Feed Forward(AFF) algorithm. The organization of this paper is as follows. In Section II, we will briefly review the Apriori method and Feed Forward Neural Network method. Section III proposes an efficient Apriori-Feed Forward(AFF) algorithm that based on Apriori and the Feed Forward(AFF) structure. Experimental results will be presented in Section IV. Section V gives out the conclusions.

II. CLASSICAL MINING ALGORITHM AND NEURAL NETWORK

A. Apriori Algorithm

In [4], Agrawal proposed an algorithm called Apriori to the problem of mining association rules first. Apriori algorithm is a bottm-up, breadth-first approach. The frequent itemsets are extended one item at a time. Its main idea is to generate k-th candidate itemsets from the (k-1)-th frequent itemsets and to

find the *k*-th frequent itemsets from the *k*-th candidate itemsets. The algorithm terminates when frequent itemsets can not be extended any more. But it has to generate a large amount of candidate itemsets and scans the data set as many times as the length of the longest frequent itemsets. Apriori algorithm can be written by pseudocode as follows.

Procedure Apriori, Input: data set D, minimum support minsup, Output: frequent itemsets L

```
(1) 1 L = \text{find\_frequent\_1\_itemsets}(D);

(2) for (k = 2; Lk \ 1 \rightarrow \phi; k++)

(3) {

(4) Ck = \text{Apriori\_gen}(Lk-1, \text{minsup});

(5) for each transactions t = D

(6) {

(7) Ct = \text{subset}(Ck, t);

(8) for each candidate c = Ct

(9) c.\text{count}++;

(10) }

(11) Lk = \{c = Ck \mid c.\text{count} > \text{minsup}\};

(12) }

(13) return L = \{L1 = L2 = ... = Ln\};
```

In the above pseudocode, Ck means k-th candidate itemsets and Lk means k-th frequent itemsets.

B. Neural Network

Neural network[19,20] is a parallel processing network which generated with simulating the image intuitive thinking of human, on the basis of the research of biological neural network, according to the features of biological neurons and neural network and by simplifying, summarizing and refining. It uses the idea of non-linear mapping, the method of parallel processing and the structure of the neural network itself to express the associated knowledge of input and output. Initially, the application of the neural network in data mining was not optimistic, and the main reasons are that the neural network has the defects of complex structure, poor interpretability and long training time. But its advantages such as high affordability to the noise data and low error rate, the continuously advancing and optimization of various network training algorithms, especially the continuously advancing and improvement of various network pruning algorithms and rules extracting algorithm, make the application of the neural network in the data mining increasingly favored by the overwhelming majority of users.

C. Neural Network Method in Data Mining

There are seven common methods and techniques of data mining[21,22,23] which are the methods of statistical analysis, rough set, covering positive and rejecting inverse cases, formula found, fuzzy method, as well as visualization technology. Here, we focus on neural network method. Neural network method is used for classification, clustering, feature mining, prediction and pattern recognition. It imitates the neurons structure of animals, bases on the M-P model and Hebb learning rule, so in essence it is a distributed matrix

structure. Through training data mining, the neural network method gradually calculates (including repeated iteration or cumulative calculation) the weights the neural network connected. The neural network model can be broadly divided into the following three types:

- (1) **Feed-forward networks**: it regards the perception back-propagation model and the function network as representatives, and mainly used in the areas such as prediction and pattern recognition;
- (2) **Feedback network**: it regards Hopfield discrete model and continuous model as representatives, and mainly used for associative memory and optimization calculation;
- (3) **Self-organization networks**: it regards adaptive resonance theory (ART) model and Kohonen model as representatives, and mainly used for cluster analysis.

D. Feedforward Neural Network:

Feedforward neural network (FF network) are the most popular and most widely used models in many practical applications. They are known by many different names, such as "multi-layer perceptrons."

Figure 2(a) illustrates a one-hidden-layer FF network with inputs $x_1, ..., x_n$ and output \hat{J} . Each arrow in the figure symbolizes a parameter in the network. The network is divided into *layers*. The input layer consists of just the inputs to the network. Then follows a *hidden layer*, which consists of any number of *neurons*, or *hidden units* placed in parallel. Each neuron performs a weighted summation of the inputs, which then passes a nonlinear *activation function*, also called the *neuron* function.

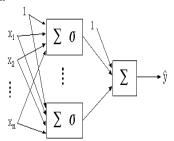


Figure 2(a) A feedforward network with one hidden layer and one output.

Mathematically the functionality of a hidden neuron is described by

$$\sigma\left(\sum_{j=1}^n w_j x_j + b_j\right)$$

where the weights $\{^{W}j, {}^{b}j\}$ are symbolized with the arrows feeding into the neuron. The network output is formed by another weighted summation of the outputs of the neurons in the hidden layer. This summation on the output is called the *output layer*. In Figure 2(a) there is only one output in the output layer since it is a single-output problem. Generally, the

number of output neurons equals the number of outputs of the approximation problem. The neurons in the hidden layer of the network in Figure 2(a) are similar in structure to those of the perceptron, with the exception that their activation functions can be any differential function. The output of this network is

$$\hat{y} (\theta) = g (\theta, x) = \sum_{i=1}^{nb} w_i^2 \sigma \left(\sum_{j=1}^{n} w_{i,j}^1 x_j + b_{j,i}^1 \right) + b^2$$

where n is the number of inputs and nh is the number of neurons in the hidden layer. The variables $\{w_{i,j}^{k_i^I}, b_{j,i}^{k_i^I}, w_{i,j}^{k_i^I}, b^2\}$ are the parameters of the network model that are represented collectively by the parameter vector 6. In general, the neural network model will be represented by the compact notation $g(\theta,x)$ whenever the exact structure of the neural network is not necessary in the context of a discussion. Note that the size of the input and output layers are defined by the number of inputs and outputs of the network and, therefore, only the number of hidden neurons has to be specified when the network is defined. The network in Figure 2(a) is sometimes referred to as a three-layer network, counting input, hidden, and output layers. However, since no processing takes place in the input layer, it is also sometimes called a two-layer network. To avoid confusion this network is called a one-hidden-layer FF network throughout this documentation.

In training the network, its parameters are adjusted incrementally until the training data satisfy the desired mapping as well as possible; that is, until $\hat{\mathcal{I}}(\bar{\mathbf{G}})$ matches the desired output y as closely as possible up to a maximum number of iterations.

The nonlinear activation function in the neuron is usually chosen to be a smooth step function. The default is the standard sigmoid

$$Sigmoid[x] = \frac{1}{1 + e^{-x}}$$

that looks like this.

<< NeuralNetworks`

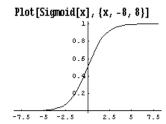


Figure 2(b)

III. APRIORI- FEEDFORWARD ALGORITHM(AFF)

In this Section, a new algorithm based on Apriori and the Feedforward Neural Network structure is presented, which is called Apriori- Feedforward Algorithm(AFF).Fig.3a shows the database structure in which there are different Item Ids and sets of Items purchased against Item ID and Fig 3(a)

Item_ID	Items
001	I_{1}, I_{2}, I_{3}
002	I ₁ , I ₃ , I ₄
003	I_2 , I_4
004	I_1, I_2
005	I ₁ , I ₂ , I ₃ , I ₅

Fig 3(a): Item Id and Item List of Database

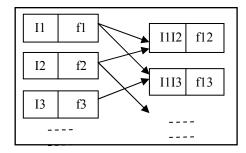


FIG 3(B): Data Structure of Nodes in FeedForward Neural Network

The Apriori- Feedforward algorithm mainly includes two

First, a neural network model is prepared according to the maximum number of items present in the dataset. Then first transaction of the data set is scanned to find out the frequent 1 itemsets, and then neural network is updated for frequent 2 itemsets frequent 3 itemsets and so on. The data set is scanned only once to build all frequent combinations of datasets. While updating frequent 2/frequent 3 itemsets..., its pruning is done at the same time to avoid redundancy of item sets. At last, the built Neural Network is mined by Apriori-FeedForward Algorithm. The detailed Apriori-FeedForward Algorithm is as follows.

Procedure: Create Model

Input: data set D, minimum support minsup

- (1) procedure Create Model(n)
- (2) for($i=1; i \neq \square; i++$)
- (3) for each itemset $l_1 \in l_{k-1}$
- (4) for each itemset $l_2 \in l_{k-1}$

(5) if(
$$l_1[1] = l_2[1]$$
) \square ($l_1[2] = l_2[2]$) \square ($l_1[3] = l_2[3]$) ($l_1[n] = l_2[n]$)

- (6) then
- (7) $C = l_1 \times l_2$
- (8) if already_generated($l_1 \times l_2$) then
- (9) delete C
- (10) else add C to C_k
- (11) $FNN(C_k)$

Procedure $FNN(C_k)$

Input: itemset Model C_k

Output: frequent itemsets L

- (1) procedure $FNN(C_k)$
- (2) n = recordcount(Dataset)
- (3) for(i=1; i < n; i++)
- (4) {
- (5) $L_1 = get first transaction(Dataset)$
- (6) $upf = update_frequecy(l_1 x l_2)$
- (7) if $(upf \ge min_sup)$
- (8) $print(l_1 \times l_2)$

In this algorithm a complete Feed forward neural network is prepared according to the maximum number of items present in the datasets. First layer of the network is a frequent 1 itemsets second layer is frequent 2 item sets and so on until a final layer is prepared which is a single node comprising of all items present in the datasets. Every n+1 layer is combination of item n with respect to all other items present at that layer, these layers are generated by calculating factorial of n+1 items.

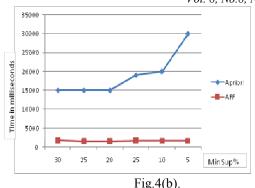
IV. EXPERIMENTAL RESULTS

The content of our test data set are frequently purchased items of a super market. There are 7 to 12 different items and 10000 to 50000 records in that data set. In order to verify the performance of the Apriori - FeedForward algorithm, we compare Apriori-Feed Forwrd with Apriori. The algorithms are performed on a computer with i7 processor 1.60GHz and 4 GB memory. The program is developed by NetBeans 6.8. The computational results of two algorithms are reported in Table 1. The clearer comparison of two algorithms is given in Fig.4(a). Table 1. The running time of two

algorithms Apriori - FeedForward algorithm

Min.Supp	Apriori	AFF
30%	15000ms	1762ms
25%	15000ms	1545ms
20%	15000ms	1529ms
15%	19000ms	1682 ms
10%	20000ms	1634 ms
5%	30000ms	1625ms

Fig.4(a). Table 1.



From Fig.4(a), 4(b). we can make the following two statements. First, Apriori- FeedForward algorithm works much faster than Apriori. It uses a different method FNN to calculate the support of candidate itemsets and it consumes less memory than Apriori because it doesn't need to traverse database again and again. It needs only single scan to the database.

V. CONCLUSION AND FUTURE WORK

In this paper, we have proposed the Apriori- FeedForward algorithm. This method builds Feed Forward Neural Network Model and scans the data base only once to generate frequent patterns. The future work is to further improve the Apriori-Feed Forward algorithm and test more and larger datasets.

REFERENCES

- [1] M.S. Chen, J. Han, P.S. Yu, "Data mining: an overview from a database perspective", *IEEE Transactions on Knowledge and Data Engineering*, 1996, 8, pp. 866-883.
- [2] J. Han, M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publisher, San Francisco, CA, USA, 2001.
- [3] R.Agrawal, T.Imielinski and A.Swami, "Mining association rules between sets of items in large databases,in: *Proceedings of the Association for Computing Machinery*, ACM-SIGMOD, 1993, 5, pp.207-216.
- [4] R. Agrawal, R. Srikant, "Fast algorithms for mining association rules", Proceedings of the 20th Very Large DataBases Conference (VLDB'94), Santiago de Chile, Chile, 1994, pp. 487-499.
- [5] Agrawal, R., Srikant, R., & Vu, Q, "Mining association rules with item constraints", In The third international conference on knowledge discovery in databases and data mining, Newport Beach, California, 1997, pp. 67-73.
- [6] J.Han, Y. Fu, "Discovery of multiple-level association rules from large database", In *The twenty-first international conference on very large* data bases, Zurich, Switzerland, 1995, pp. 420-431.
- [7] Fukuda, T., Morimoto, Y., Morishita, S., & Tokuyama, T., "Mining optimized association rules for numeric attributes", In *The ACM SIGACT-SIGMOD-SIGART symposium on principles of database systems*, 1996, pp. 182-191.
- [8] Park, J. S., Chen, M. S., & Yu, P. S., "Using a hash-based method with transaction trimming for mining association rules", *IEEE Transactions* on Knowledge and Data Engineering, 1997, 9(5), pp. 812-825.
- [9] J.Han, J.Pei and Y.Yin., "Mining frequent patterns without candidate Generation", in: *Proceeding of ACM SIGMOD International Conference Management of Data*, 2000, pp. 1-12.

Vol. 8, No.8, November 2010

- [10] J.Han, J.Wang, Y.Lu and P.Tzvetkov, "Mining top-k frequent closed patterns without minimum support", in: Preceeding of International Conference Data Mining, 2002, 12, pp. 211-218.
- [11] G.Liu, H.Lu, J.X.Yu, W.Wei and X.Xiao, "AFOPT: An efficient implementation of pattern growth approach", in: IEEE ICDM Workshop Frequent Itemset Mining Implementations, CEUR Workshop Proc., 2003, 80.
- [12] J.Wang, J.Han, and J.Pei, "CLOSET+: searching for the best strategies for mining frequent closed Itemsets", in: Preceeding of International Conference, Knowledge Discovery and Data Mining, 2003, 8, pp. 236-245
- [13] Tzung-Pei Hong, Chun-Wei Lin, Yu-Lung Wu, "Incrementally fast updated frequent pattern trees", Expert Systems with Applications, 2008, 34, pp. 2424-2435.
- [14] K. Wang, Y. He, D. Cheung, Y. Chin, "Mining confident rules without support requirement", in: *Proceedings of ACM International Conference* on Information and Knowledge Management, CIKM, 2001, pp.89-96.
- [15] H. Xiong, P. Tan, V. Kumar, "Mining strong affinity association patterns in data sets with skewed support distribution", in: *Proceedings of the Third IEEE International Conference on Data Mining*, ICDM, 2003, pp. 387-394.
- [16] Ya-Han Hu, Yen-Liang Chen, "Mining association rules with multiple minimum supports: a new mining algorithm and a support tuning mechanism", *Decision Support Systems*, 2006, 42, pp. 1-24.
- [17] J. Ding, "Efficient association rule mining among infrequent items", *Ph.D. Thesis*, University of Illinois at Chicago, 2005.
- [18] Ling Zhou, Stephen Yau, "Efficient association rule mining among both frequent and infrequent items", Computers and Mathematics with Applications, 2007, 54, pp. 737-749.

- [19] Anderson, J. A., 1995, Introduction to Neural Networks (Cambridge, MA:MIT Press).
- [20] Van Hulle, M. M., 2000, Faithful Representations and Topographic Maps:From Distortion-to-Information-Based Self Organization (New York:Wiley).
- [21] Cristofor, L., Simovici, D., Generating an informative cover for association rules. In Proc. of the IEEE International Conference on Data Mining, 2002.
- [22] Yuan, Y., Huang, T., A Matrix Algorithm for Mining Association Rules, Lecture Notes in Computer Science, Volume 3644, Sep 2005, Pages 370
- [23] Sotiris Kotsiantis, Dimitris Kanellopoulos, Association Rules Mining: A Recent Overview, GESTS International Transactions on Computer Science and Engineering, Vol.32 (1), 2006, pp. 71-82

An Efficient Vector Quantization Method for Image Compression with Codebook generation using Modified K-Means

S. Sathappan, Associate Professor of Computer Science, Erode Arts and Science College, Erode-638 009.

Tamil Nadu. India.

Abstract— With the growth of internet and multimedia, compression techniques have become the thrust area in the fields of computers. Image compression is a technique of efficiently coding digital image to reduce the number of bits required in representing image. Many image compression techniques presently exist for the compression of different types of images. In this paper Vector Quantization based compression scheme is introduced. In this scheme a low bit rate still image compression is performed by compressing the indices of Vector Quantization and residual codebook is generated. The indices of VQ are compressed by exploiting correlation among image blocks, which reduces the bit per index. A residual codebook similar to VQ codebook is generated that represents the distortion produced in VQ. Using this residual codebook the distortion in the reconstructed image is removed, thereby increasing the image quality. The proposed technique combines these two methods and by replacing the Modified k-means algorithm for LBG in the codebook generation. Experimental results on standard image Lena show that the proposed scheme can give a reconstructed image with a higher PSNR value than all the existing image compression techniques.

Keywords—Image compression, Vector Quantization, Residual Codebook, Modified K-Means

I. INTRODUCTION

MAGE compression is a method of efficiently coding digital image, to reduce the number of bits required in representing image. Its purpose is to decrease the storage space and transmission cost while maintaining good quality. VECTOR Quantization [1] has been found to be an efficient technique for image compression in the past decade. VQ compression system mainly contains two components: VQ encoder and decoder as shown in Fig.1.

In VQ technique [2] [3], the input image is partitioned into a set of non-overlapping image blocks $X = \{x_0, x_1, \dots, x_{m-1}\}$

of size 4x4 pixels each and a clustering algorithm, for example Linde–Buzo–Gray (LGB) algorithm [5] and Modified K-Means [2]. The Modified K-Means algorithm is used in the proposed technique, to generate a codebook $C = \{V_0, V_1, \dots, V_{N-1}\}$ for the given set of image blocks. The

codebook C comprises a set of representative image blocks called codewords. The VQ encoder discovers a closest match codeword in the codebook for each of the image block and the index of the codeword is transmitted to VQ decoder. The decoder phase has the following functionalities. VQ decoder replaces the index values with the respective codewords from

the codebook and produces the quantized image, called as reconstructed image. In order to attain low bit rate, many VQ schemes, have been used in the past literature such as sidematch VQ (SMVQ) [6], classified SMVQ (CSMVQ) [7] and Gradient based SMVQ (GSMVQ) [8].

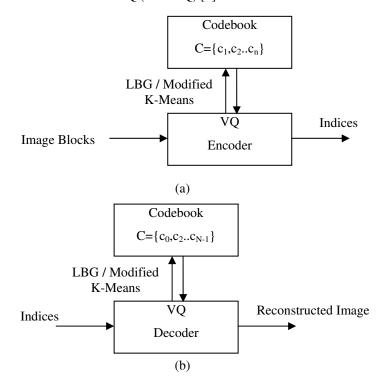


Fig. 1 (a) VQ Encoder (b) VQ Decoder

Even though, SMVQ, CSMVQ, GSMVQ and JSPVQ are the low bit rate schemes, they require high encoding time than that of VQ method. In this paper, an efficient low bit rate image compression scheme is proposed based on VQ that makes use of compression of indices of VQ and residual codebook with modified k-means clustering instead of LGB. This scheme attains low bit rate and better image quality than previous VQ methods.

The rest of the paper is presented as follows: in section II, the literature survey is presented. In section III the proposed compression scheme is described. Performance of the proposed system is evaluated in section IV and section V concludes the paper.

II. LITERATURE SURVEY

Somasundaram et al, [1] presented a novel vector quantization approach and also explained various VQ schemes, such as side-match VQ (SMVQ), classified SMVQ (CSMVQ) and Gradient based SMVQ (GSMVQ). In which SMVQ uses the high correlation existing between neighboring blocks to achieve low bit rate and master codebook C is used to encode image blocks in the first column and first row in advance. The other image blocks are encoded, utilizing the correlation with the neighboring encoded image blocks. Let x be the input image block for the compression system, and u and 1 be the upper and left neighboring codewords respectively. Let the size of the given image block size be $k = m \times n$. The side-match distortion of a codeword Y can be defined as:

$$smd(Y) = \sum_{i=0}^{n-2} (u_{(m-1,i)} - Y_{(0,i)})^2 + \sum_{i=0}^{m-1} (l_{(i,n-1)} - Y_{(i,0)})^2$$
(1)

According to their side-match distortions of all codewords SMVQ sorts the codewords and then selects N_S codewords with smallest side-match distortions from the master book C of size N to form the state codebook SC, where $N_S < N$. A best-match codeword Y_i is selected to encode an image block x from N_S codewords and the corresponding index is coded in log₂N_S bits. Thus, the SMVQ reduces the bit rate of VQ. Since mean square error caused by state codebook is higher than that of master codebook, SMVQ degrades the image quality and also it requires long encoding time. Classified side-match vector quantization [7] (CSMVQ) is an efficient low bit rate image compression technique which produces relatively high quality image. It is a variable rate SMVO and makes use of variable sized state codebooks to encode the current image block. The size of the state codebook is decided based on the variances of left codewords and upper codewords that predict the block activity of the input blocks. Also, CSMVQ uses two master codebooks, one for low detail blocks and another for high detail blocks. Another variant, gradient-based SMVQ [8] (GSMVQ) has been proposed, in which gradient values are used instead of variance values to predict the input vector. Another low bit rate VQ, called Jigsaw-puzzle vector quantization (JPVQ) [9] was proposed, in which an input block can be coded by the super codebook, the dynamic codebook or the jigsaw-puzzle block. The jigsawpuzzle block is constructed dynamically using four-step side match prediction technique.

Interpolative vector quantization, first proposed explicitly by Gersho in [12], introduces dimension reduction to traditional VQ. The codebook in the encoder is learned on downsampled vectors and the codebook in the decoder on high-dimension vectors. Except for the difference on dimension, the two codebooks have the same number of representative vectors and structure. VQ encoder maps down the sampled inputs to a set of scalar indices and VQ decoder reproduces high-dimension inputs by received indices. David

et al. applied IVQ to image restoration [13], where the encoder does not need a codebook except for some parameters. The codebook at decoder is learned on image pairs consisting of an original image and its diffraction-limited counterpart. Several follow-up work is reported in [14][15].

The goal of quantization is to encode the data from a source, with some loss, so that the best reproduction is obtained. Vector quantization (VQ) achieves compression then scalar quantization [14], making it useful for band-limited channels. The algorithm for the design of optimal VQ is commonly referred to as the Linde-Buzo-Gray (LBG) algorithm, and it is based on minimization of the squared-error distortion measure. The LBG algorithm starts with an initial codebook and iteratively partitions the training sequence into the Voronoi regions to obtain a new codebook that produces a lower distortion. Once the final codebook is got, it can be used on new data outside the training sequence with the optimum nearest neighbor rule. If the training sequence is sufficiently long, it yields good performance for future data produced by the source.

In the paper by M.Antonini, et al. [18], images have been coded using two-level wavelet decomposition with VQ of the resulting coefficients. A multiresolution codebook has been designed with noise-shaping bit allocation among the various subbands. The test image has been coded at a rate of 0.78bpp, achieving a PSNR of 32.1dB. In the paper by Gersho and Ramamurthy [19], images have been compressed using unstructured VQ, achieving a bitrate of 0.5 – 1.5bpp. Ho and Gersho [19] have used multistage VQ for progressive image coding, with a PSNR of 30.93dB at 0.36bpp using 4 stages. R.L.Claypoole et al. [21] have coded images using nonlinear wavelet transform via lifting and obtained 30.5dB at 0.6bpp. An adaptive lifting scheme with perfect reconstruction has been used in [21].

III. METHODOLOGY

The compression scheme consists of two components, compression of indices and generation of residual codebook. These two are explained in this section.

3.1. Compression of Indices

The Index compression step has the following procedure. When the image blocks are vector quantized, there likely to exist high correlation among the neighboring blocks and hence among the corresponding codeword indices. Therefore, if indices are coded by comparing with the previous indices, further reduction in the bit rate can be achieved. In Search Order Coding (SOC) [11], a simple searching scheme is followed to find a match for the current index from the previous indices. The search order SO is defined as the order in which the current index is compared with the previous indices.

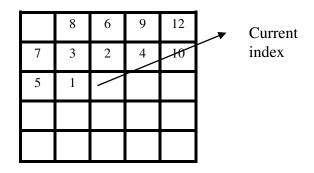


Fig. 2 Searching Order

The SO used in [11] is given in Fig.2. The label "1" indicates the highest searching priority, "2" denotes the second highest priority and so on. In order to limit comparisons of the current index with previous indices, the searching range (SR) is fixed. The SR is defined as the number of previous indices to be compared with current index. In this paper Search Order [11], SR is taken as 10, which gives the lower bit rate than other SR.

In this method, the index of the codeword of a block is encoded exploiting the degree of the similarity of the block with previously encoded upper or left blocks. When the degree of similarity of the current block with one of the two previously encoded blocks is high, the index of the codeword of the block is encoded using the index of the neighboring codeword. I.e. the codeword index of the current block and that of the neighboring blocks are same. If the degree of similarity of the block with the neighboring blocks is not high, it is assumed that the closest match codeword of the current block may be nearer to the codewords of the neighboring blocks. For example, if one of the two neighboring blocks codeword's index is 'N', the closest match codeword of the block to be encoded may lie between (N-J) th codeword and (N+J) th codeword in the codebook, where J is any arbitrary number. So the index can be coded in log₂(2*J) bits. This idea is based on the property existing in the codebook design using LBG algorithm with splitting technique. In the splitting technique, bigger size codebook is generated by splitting each codeword of the smaller codebook into two. The size of the codebook is always in powers of two $(2^{M} \rightarrow 2^{(M+1)})$. Hence, relatively similar two image blocks may have same closest match codeword in Jth position at codebook of size 2^M and at codebook of size 2^(M+1), one of the two image blocks may have its closest match codeword at Jth place in the codebook and other block's codeword may be in (J+1)th place. The other non-similar blocks are encoded using their original index value. In this scheme, examining the resemblance of a block with its left and upper blocks is not required to encode the index of the block. The above description is the idea behind our VO indices compression scheme. In order to implement this idea, the index to be coded is compared with previous indices according to the SO given in Fig.2 and SR is fixed as 2 in this scheme. Let 1, 2,..,12 be the SO and ind_val (1), ind_val (2),..ind_val(12) be the indices values of the SO = 1,2,...12. The following steps is used to encode VQ index.

- 1 Get the first index generated by the VQ encoder and transmit as such.
- 2. Get the next index generated by VQ Encoder. Compare this index with the previous indices according SO
- 3. if SO = 1, code it as "00" and go to the step 2 else if SO = 2, code it as "01" and go to the step 2
- else go to the next step. 4 if index value \leq (ind_val (SO = 1) + J) and index value \geq — (ind_val (SO = 1)+J) { if ind_val (SO = 1) = ind_val (SO=2) code it as "10" followed by log (2 *) 2 J bits else

code it as "100" followed by log (2 *) 2 J bits. }

go to step 2.

else

if index value \leq (ind_val (SO = 2) + J) and index value \geq — (ind_val (SO = 2)+J) code it as "101" followed by log (2 *) 2 J bits and go to step 2.

else

code it as "11" followed by its original index and goto step 2.

Decoding of the compressed indices is done by reversing the above coding steps.

3.2. Construction of Residual Codebook (RC)

The residual codebook can be represented as $RC = \{RY_0, RY_1, \dots, RY_{k-1}\}$. Residual codebook is

constructed using absolute error values caused by VQ method, in the residual codebook construction, the image blocks that are less similar to their closest match codewords found in the codebook are taken into account. Less similarity blocks will increase distortion than high similarity blocks in the reconstructed image. Residual codeword (RY_i) for a less similarity image block is constructed by comparing it with its closest match codeword. The collection of residual codewords RY_i, RY_{i+1}... is called residual codebook. Similarity of an image block x with its closet match codeword Y_i is determined based on minimum distortion rule (α) between them. If the mean square error (α) of an image block is greater than a predefined threshold value (α), then the block is taken as less similarity block.

Let $x = (x_0, x_1, \dots, x_{k-1})$ be a k-pixels image block and $Y_t = (y_0, y_1, \dots, y_k)$ be a k-pixels closest match codeword,

then the α is defined as:

$$\alpha = \frac{1}{k} \sum_{i=0}^{k-1} (x_i - y_i)^2$$
(2)

The steps used for constructing residual codebook are given below.

Step 1: An image which is to be compressed is decomposed into a set of non-overlapped image blocks of 4x4 pixels.

Step 2: A codebook is generated for the image blocks using LBG algorithm.

Step 3: Pick up the next codeword Y_t from the codebook C and find its all closest match less similarity image blocks(X) found out using (2) from the given set of image blocks and construct residual codeword RY_t using the following equation.

$$RY_{r} = \frac{1}{m} \sum_{i=1}^{m} \{ |Y_{r1} - X_{i1}|, |Y_{r2} - X_{i2}|, \dots, |Y_{rk} - X_{ik}| \}$$
(3)

where k represents the number of elements in the codeword Y_t and the image block Xi respectively and m denotes the number of less similarity image blocks that are closer to the codeword Y_t .

Repeat the step 3 until no more codeword exists in the codebook. Since residual codeword RY_i is constructed only for less similarity image blocks, some of the codewords Y_i may not have their respective residual codewords, i.e; these codewords may not have less similarity image blocks. In residual codebook construction, only absolute values of the residuals of the less similarity image blocks are used. The sign information for each less similarity image block is preserved and is called residual sign bit plane. In encoding phase, for each less similarity image block, pixels of the block are subtracted from the corresponding pixel values of the codeword Yi, then sign values (positive or negative) of the residual values of that block, called residual sign bit plane, are preserved. To reduce the bits needed for residual sign bit plane, only alternate bits are stored and others are dropped based on the assumption that there exists correlation among neighboring bits. The bits used for prediction is shown in Fig.3. In the decoding process, the bits of the residual sign bit plane of a block are replaced with the respective residual values of the residual codeword from the residual codebook (RC) with appropriate sign. The residual values of the dropped bits are predicted from neighboring residual values using following steps.

$$\begin{aligned} 1.pv(B) &= \frac{rrv(A) + rrv(C) + rrv(F)}{3} \\ 2.pv(D) &= \frac{rrv(C) + rrv(H)}{2} \\ 3.pv(E) &= \frac{rrv(A) + rrv(I) + rrv(F)}{3} \\ 4.pv(G) &= \frac{rrv(H) + rrv(C) + rrv(F) + rrv(K)}{4} \\ 5.pv(J) &= \frac{rrv(I) + rrv(N) + rrv(F) + rrv(K)}{4} \\ 6.pv(L) &= \frac{rrv(H) + rrv(K) + rrv(F)}{3} \end{aligned}$$

$$7.pv(M) = \frac{rrv(I) + rrv(N)}{2}$$

$$8.pv(O) = \frac{rrv(N) + rrv(K) + rrv(P)}{3}$$

where pv (*) is the pre assigned value of the corresponding bit in the residual sign bit plane and rrv (*) is the respective reconstructed residual value of the bit in the residual sign bit plane.

Fig. 3 Bits encircled are used for prediction

After reconstructing the residual codeword, each value of the residual codeword is added to respective value of the closest match codeword of the block Since the residual sign bit plane for each image block has only eight bits, alternate residual values in the residual codeword RY_t are dropped and it also reduces the cost of storing residual codebook. The dropped residual values are predicted from the neighboring residual values as given above.

3.3. Modified K-Means Algorithm to Replace LBG

Initial Cluster Centers Deriving from Data Partitioning

The algorithm follows a novel approach that performs data partitioning along the data axis with the highest variance. The approach has been used successfully for color quantization [9]. The data partitioning tries to divide data space into small cells or clusters where intercluster distances are large as possible and intracluster distances are small as possible.

For instance, Suppose ten data points in 2D data space are given. The goal is to partition the ten data points into two disjoint cells where sum of the total clustering errors of the two cells is minimal. Suppose a cutting plane perpendicular to X-axis will be used to partition the data. Let C_1 and C_2 be the first cell and the second cell respectively and $\overline{C_1}$ and $\overline{C_2}$ be the

cell centroids of the first cell and the second cell, respectively. The total clustering error of the first cell is thus computed by:

$$\sum_{i \in \mathcal{C}_i} d\left(c_i, \bar{c_1}\right) \tag{4}$$

and the total clustering error of the second cell is thus computed by:

$$\sum_{c_1 \in C_2} d(c_1, \overline{c_2}) \tag{5}$$

Where c_i is the ith data in a cell. As a result, the sum of total clustering errors of both cells is minimal.

The partition could be done using a cutting plane that passes through m. Thus

$$d(e_i, \overline{e_1}) \le d(e_i, e_m) + d(\overline{e_1}, e_m) \text{ and}$$

$$d(e_i, \overline{e_2}) \le d(e_i, e_m) + d(\overline{e_2}, e_m)$$
(6)

Thus
$$\sum_{c_i \in C_1} d(c_i, \overline{c_1}) \le \sum_{c_i \in C_2} d(c_i, c_m) + d(\overline{c_2}, c_m), |C_2|$$

$$\sum_{c_i \in C_2} d(c_i, \overline{c_2}) \le \sum_{c_i \in C_2} d(c_i, c_m) + d(\overline{c_2}, c_m), |C_2| \tag{7}$$

m is called as the partitioning data point where |C1| and |C2| are the numbers of data points in cluster C1 and C2 respectively. The total clustering error of the first cell can be minimized by reducing the total discrepancies between all data in first cell to m, which is computed by:

$$\sum_{c_i \in C_i} d(c_i, c_m) \tag{8}$$

The same argument is also true for the second cell. The total clustering error of second cell can be minimized by reducing the total discrepancies between all data in second cell to m, which is computed by:

$$\sum_{c_i \in C_0} d\left(c_i, c_m\right) \tag{9}$$

where $d(c_i,c_m)$ is the distance between m and each data in each cell. Therefore the problem to minimize the sum of total clustering errors of both cells can be transformed into the problem to minimize the sum of total clustering error of all data in the two cells to m.

The relationship between the total clustering error and the clustering point is illustrated in Fig. 4, where the horizontal-axis represents the partitioning point that runs from 1 to n where n is the total number of data points and the vertical-axis represents the total clustering error. When m=0, the total clustering error of second cell equals to the total clustering error of all data points while the total clustering error of first cell is zero. On the other hand, when m=n, the total clustering error of all data points, while the total clustering error of all data points, while the total clustering error of the second cell is zero.

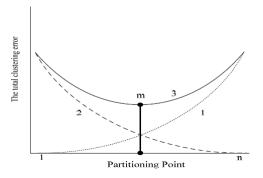


Fig. 4 Graphs depict the total clustering error, lines 1 and 2 represent the total clustering error of the first cell and second cell, respectively, Line 3 represents a summation of the total clustering errors of the first and the second cells

A parabola curve shown in Fig. 4 represents a summation of the total clustering error of the first cell and the second cell, represented by the dash line 2. Note that the lowest point of the parabola curve is the optimal clustering point (m). At this point, the summation of total clustering error of the first cell and the second cell are minimum.

Since time complexity of locating the optimal point m is $O(n^2)$, the distances between adjacent data is used along the X-axis to find the approximated point of n but with time of O(n).

Let
$$D_j = d(c_j, c_{j-1})^2$$
 be the squared Euclidean distance of adjacent data points along the X-axis.

If i is in the first cell then $d(e_m, e_i) \leq \sum_{j=1}^m D_j$. On the one

hand, if i is in the second cell then
$$d(e_m, e_i) \leq \sum_{j=m}^m D_j$$

The task of approximating the optimal point (m) in 2D is thus replaced by finding m in one-dimensional line.

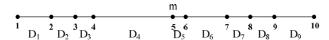


Fig. 5 Illustration of the ten data points on a one-dimensional line and the relevant Dj

The point (m) is therefore a centroid on the one dimensional line (as shown in Fig. 5), which yields

$$\sum_{i=1}^{m-1} d(e_m, e_i) \approx \sum_{i=m}^{n} d(e_m, e_i)$$
(10)

Let $dsum_i = \sum_{j=1}^{i} D_j$ and a *centroidDist* can be computed

$$centroidDist = \frac{\sum_{i=1}^{n} dsum_i}{n}$$
 (11)

It is probable to choose either the X-axis or Y-axis as the principal axis for data partitioning. However, data axis with the highest variance will be chosen as the principal axis for data partitioning. The reason is to make the inter distance between the centers of the two cells as large as possible while the sum of total clustering errors of the two cells are reduced from that of the original cell. To partition the given data into k cells, it is started with a cell containing all given data and partition the cell into two cells. Later on the next cell is selected to be partitioned that yields the largest reduction of total clustering errors (or Delta clustering error). This can be described as Total clustering error of the original cell - the sum of Total clustering errors of the two sub cells of the original cell. This is done so that every time a partition on a cell is performed, the partition will help reduce the sum of total clustering errors for all cells, as much as possible.

The partitioning algorithm can be used now to partition a given set of data into k cells. The centers of the cells can then be used as good initial cluster centers for the K-means algorithm. Following are the steps of the initial centroid predicting algorithm.

- 1. Let cell c contain the entire data set.
- 2. Sort all data in the cell c in ascending order on each attribute value and links data by a linked list for each attribute.
- 3. Compute variance of each attribute of cell c. Choose an attribute axis with the highest variance as the principal axis for partitioning.
- 4. Compute squared Euclidean distances between adjacent data along the data axis with the highest variance $D_i = d(e_i, e_{i+1})^2$ and compute the $dsum_i = \sum_{j=1}^i D_j$
 - 5. Compute centroid distance of cell c:

$$centroidDist = \frac{\sum_{i=1}^{n} dsum_i}{n}$$

Where $dsum_i$ is the summation of distances between the adjacent data.

- 6. Divide cell c into two smaller cells. The partition boundary is the plane perpendicular to the principal axis and passes through a point m whose dsumi approximately equals to centroidDist. The sorted linked lists of cell c are scanned and divided into two for the two smaller cells accordingly
- 7. Calculate Delta clustering error for c as the total clustering error before partition minus total clustering error of its two sub cells and insert the cell into an empty Max heap with Delta clustering error as a key.
- 8. Delete a max cell from Max heap and assign it as a current cell.
- 9. For each of the two sub cells of c, which is not empty, perform step 3 7 on the sub cell.
- 10. Repeat steps 8 9. Until the number of cells (Size of heap) reaches K.
- 11. Use centroids of cells in max heap as the initial cluster centers for K-means clustering

3.4. The Proposed Algorithm

The proposed scheme combines compression of VQ indices and residual codebook. The steps used in this compressor are as follows

- 1. An image which is to be compressed is decomposed into a set of non-overlapped image blocks of size 4x4 pixels.
- 2. A codebook is generated for the image blocks using Modified K-Means algorithm.
- 3. Construct a Residual Codebook (as described in section 3.4) for those image blocks (less similarity blocks) whose α is greater than σ .
- 4. Pick the next image block (current block) and find its closest match codeword in the codebook. Calculate mean square error α for the image block using equation (2) and index of the codeword is encoded using VQ indices compression scheme presented in section 3.1.

5. if $(\alpha \le \sigma)$, the current block is encoded as "0". else

the current block is encoded as "1" followed by interpolated residual sign bitplane which is computed as described in section 3.2.

6. Repeat the step 4 until no more blocks exist in the image.

The decoding of the compressed images is done by reversing the above said steps and residual block to be added is reconstructed for each less similarity block as described in section 3.2.

IV. EXPERIMENTAL RESULTS

To evaluate the proposed technique experiments are carried out on standard gray scale images using a Pentium-IV computer running at 1.60 GHz under Windows XP. Three images of 512 x 512 pixels in size are used. Codebook is generated using Modified K-Means algorithm for all the methods. Codebook is also generated with LBG [5] for comparison. For this scheme, a codebook of size 64 is used. Performances of the above algorithms are evaluated in terms of bit rate (bits per pixel) and peak signal-to-noise ratio (PSNR) given by:

$$PSNR = 10 \log_{10} \frac{(255)^2}{MSE} db$$

where MSE (mean squared error) is defined as:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (x_i - y_i)$$

where x_i and y_i denote the original and the encoded pixel values and n is the total number of pixels in an image. Bit rate including overhead bits (i.e bits need to store codebook) for different threshold values ranging from 50 to 2000 for Lena, Camera man and Pepper.

The performance of proposed scheme is evaluated with the existing techniques for different gray-scale images of size 512x512 and is given in the table I. J is set to 4 for the proposed scheme. From table I, it can be observed note that proposed method with the modified k-means instead of LBG has an improvement in coding the VQ indices.

TABLE I

PERFORMANCE OF PROPOSED METHOD VQ WITH CODEBOOK SIZE 64 USING LBG, K-MEANS AND MODIFIED K-MEANS IN CODING STANDARD GRAY SCALE IMAGES OF SIZE 512 x 512

EACH Modified vo **LBG** K-Means **Images** K-Means bits/index (bits/index) (bits/index) (bits/index) 3.92 3.88 Lena 6 3.47 4.08 3.99 3.32 Cameraman 6 3.72 3.66 3.20 **Peppers**

Table II shows the comparison of the PSNR values for the Lena, Camera man and Pepper images of 512x512 bits when compressed with combine VQ method and Codebook using LBG, K-Means and Modified k-Means algorithm.

 $\begin{array}{c} \text{Table II} \\ \text{Comparison of PSNR values for three standard} \\ \text{IMAGES} \end{array}$

Images	VQ bits/index	LBG (dB)	K- Means (dB)	Modified K-Means (dB)	
Lena	6	31.60	32.51	34.65	
Cameraman	6	30.24	31.25	33.84	
Peppers	6	31.44	32.26	34.20	

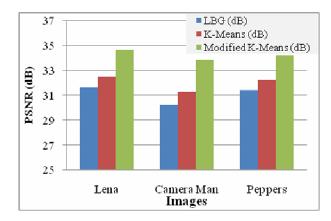


Fig. 6 Comparison of PSNR values for three standard images From Fig. 6 it is observed that this scheme gives PSNR values of 34.65db, 33.84db and 34.20 db for Lena, Camera man and Peppers respectively. From this it can be observed that the proposed approach produces better result than all the existing methods.

V. CONCLUSION

The rapid increase in the range and use of electronic imaging justifies attention for systematic design of an image compression system and for providing the image quality needed in different applications. There are a lot of techniques available for image compression. In this paper, a new gray scale image compression scheme is proposed which gives better image quality and low bit rate. This scheme is based on VQ method and employs residual codebook to improve image quality and compression of VQ indices to lower the bit rate. Experimental results on standard images show that the proposed scheme gives better PSNR values and low bit rate than previous methods with codebook generation using LBG and kmeans. Since this scheme uses smaller codebook, it gives faster compression than the other two schemes.

REFERENCES

- [1] K.Somasundaram, and S.Domnic, "Modified Vector Quantization Method for Image Compression", World Academy of Science, Engineering and Technology 19 2006, pp:128-134
- [2] S. Deelers, and S. Auwatanamongkol, "Enhancing K-Means Algorithm with Initial Cluster Centers Derived from Data Partitioning along the Data Axis with the Highest Variance", International Journal of Computer Science Volume 2 Number 4, pp:247-252
- [3] R. M. Gray, "Vector quantization," IEEE Acoustics, speech and Signal Processing Magazine, pp. 4-29, 1984.
- [4] M. Goldberg, P. R. Boucher and S. Shlien, "Image Compression using adaptive vector quantization," IEEE Transactions on Communication, Vol. 34, No. 2, pp. 180-187, 1986.
- [5] Y. Linde, A. Buzo and R. M. Gray, "An algorithm for vector quantizer design," IEEE Transactions on Communication, Vol. 28, No. 1, 1980, pp. 84 – 95.
- [6] T.Kim, "Side match and overlap match vector quantizers for images," IEEE Trans. Image. Process., vol.28 (1), pp.84-95, 1980.
- [7] Z.M.Lu, J.S Pan and S.H Sun, "Image Coding Based on classified sidematch vector quantization," IEICE Trans.Inf.&Sys., vol.E83-D(12), pp.2189-2192, Dec. 2000.
- [8] Z.M.Lu, B.Yang and S.H Sun, "Image Compression Algorithms based on side-match vector quantizer with Gradient-Based classifiers," IEICE Trans.Inf.&Sys., vol.E85-D(9), pp.1414-1420, September. 2002.
- [9] Chia-Hung Yeh, "Jigsaw-puzzle vector quantization for image compression", Opt.Eng Vol.43, No.2, pp. 363-370, Feb-2004.
- [10] C.H.Hsieh, and J.C Tsai, "Lossless compression of VQ index with search order Coding," IEEE Trans. Image Processing, Vol.5, No. 11, pp. 1579-1582, Nov. 1996.
- [11] Chun-Yang Ho, Chaur-Heh Hsieh and Chung-Woei Chao, "Modified Search Order Coding for Vector Quantization Indexes," Tamkang Journal of Science and Engineering, Vol.2, No.3, pp. 143-148, 1999.
- [12] Gersho, "Optimal nonlinear interpolative vector quantization", IEEE trans. on communication, vol. 38, pp 1285-1287, 1990.
- [13] D. G. Sheppard, A. Bilgin, M. S. Nadar, B. R. Hunt, M. W. Marcellin, "A vector quantization for image restoration", IEEE trans. on Image Processing, vol. 7, pp119-124, 1998.
- [14] R. Nakagaki, A. K. Katsaggelos, "A VQ-based blind image restoration algorithm", IEEE trans. on Image Processing, vol. 12, pp 1044-1053, 2003
- [15] Y. C. Liaw, W. Lo, Z. C. Lai, "Image restoration of compressed image using classified vector quantization", Pattern Recognition, vol. 35, pp329-340, 2002.
- [16] M.A. Cody, The Fast Wavelet Transform, Dr. Dobb's Journal, pp. 16-28, April 1992.
- [17] R.C. Gonzalez, R.E. Woods, Digital Image Processing, Pearson Education Pvt. Ltd., New Delhi, 2nd Edition.
- [18] M. Antonini, M. Barlaud, P. Mathieu, I. Daubechies, Image Coding using Wavelet Transform, IEEE Transactions on Image Processing, Vol. 1, No. 2, pp. 205-220, April, 1992.
- [19] A.Gersho, B.Ramamurthy, Image Coding Using Vector Quantization, Proceedings of IEEE International Conference On Acoustics, Speech and Signal Processing, pp. 428-431, May 1982.
- [20] Y.Ho, A.Gersho, Variable-Rate Multistage VQ for Image Coding, Proceedings of IEEE International Conference On Acoustics, Speech and Signal Processing, pp.1156-1159, 1988.
- [21] R.L.Claypoole, Jr., G.M.Davis, W.Sweldens, R.G.Baraniuk, Nonlinear Wavelet Transforms for Image Coding via Lifting, IEEE Transactions on Image Processing, Vol. 12, No.12, pp. 1449 –1459, Dec. 2003.

Optimization of work flow execution in ETL using Secure Genetic Algorithm

Raman Kumar¹, Saumya Singla², Sagar Bhalla³ and Harshit Arora ⁴
^{1,2,3,4} Department of Computer Science and Engineering,
^{1,2,3,4} D A V Institute of Engineering and Technology, Jalandhar, Punjab, India.

Abstract— Data Warehouses (DW) typically grows asynchronously, fed by a variety of sources which all serve a different purpose resulting in, for example, different reference data. ETL is a key process to bring heterogeneous and asynchronous source extracts to a homogeneous environment. The range of data values or data quality in an operational system may exceed the expectations of designers at the time validation and transformation rules are specified. Data profiling of a source during data analysis is recommended to identify the data conditions that will need to be managed by transformation rules and its specifications. This will lead to implementation of the ETL process. Extraction-Transformation-Loading (ETL) tools are set of processes by which data is extracted from numerous databases, applications and systems transformed as appropriate and loaded into target systems - including, but not limited to, data warehouses, data marts, analytical applications, etc. Usually ETL activity must be completed in certain time frame. So there is a need to optimize the ETL process. A data warehouse (DW) contains multiple views accessed by queries. One of the most important decisions in designing a data warehouse is selecting views to materialize for the purpose of efficiently supporting decision making. Therefore heuristics have been used to search for an optimal solution. Evolutionary algorithms for materialized view selection based on multiple global processing plans for queries are also implemented. The ETL systems work on the theory of random numbers, this research paper relates that the optimal solution for ETL systems can be reached in fewer stages using genetic algorithm. This early reaching of the optimal solution results in saving of the bandwidth and CPU time which it can efficiently use to do some other task. Therefore, the proposed scheme is secure and efficient against notorious conspiracy goals, information processing.

Keywords- Extract, Transform, Load, Data Warehouse(DW), Genetic Algorithm (GA), Architecture, Information Management System, Virtual Storage Acess Method and Indexed Sequential Access Method

I. INTRODUCTION

Companies know they have valuable data lying around throughout their networks that needs to be moved from one place to another—such as from one business application to another or to a data warehouse for analysis. The only problem is that the data lies in all sorts of heterogeneous systems and therefore in all sorts of formats. To integrate data to one warehouse for analysis a tool is required which can integrate data from various systems. To solve the problem, companies use extract, transform and load (ETL) software. Usually ETL

activity must be completed in certain time frame. So there is a need to optimize the ETL process. Typical ETL activity consists of three major tasks: extraction, transformation and loading.

This research paper is the study of extraction and transformation stages. Data extraction can be seen as reader writer problem, which has been reformulated using multiple buffers instead of using single finite buffer. Transformation is set of activities which convert the data from one form to other. This thesis studies the use of Genetic Algorithm to optimize the ETL workflow.

A. Basic concepts of ETL

a) Extract

The first part of an ETL process involves extracting the data from the source systems. Most data warehousing projects consolidate data from different source systems. Each separate system may also use a different data organization/format. Common data source formats are relational databases and flat files, but may include non-relational database structures such as Information Management System (IMS) or other data structures such as Virtual Storage Access Method (VSAM) or Indexed Sequential Access Method (ISAM), or even fetching from outside sources such as web speeding or screen-scraping. Extraction converts the data into a format for transformation processing. An intrinsic part of the extraction involves the parsing of extracted data, resulting in a check if the data meets an expected pattern or structure. If not, the data may be rejected entirely.

b) Transform

The transform stage applies to a series of rules or functions to the extracted data from the source to derive the data for loading into the end target. Some data sources will require very little or even no manipulation of data. In other cases, one or more of the following transformations types to meet the business and technical needs of the end target may be required:

- Selecting only certain columns to load (or selecting null columns not to load).
- Translating coded values (For example, if the source system stores 1 for male and 2 for female, but the warehouse stores M for male and F for female), this

calls for automated data cleansing; no manual cleansing occurs during ETL.

- Encoding free-form values (For example, mapping "Male" to "1" and "Mr" to M).
- Deriving a new calculated value (For example, sale_amount = qty * unit_price).
- Filtering.
- Sorting.
- Joining data from multiple sources (For example, lookup, merge).
- Aggregation (for example, rollup summarizing multiple rows of data total sales for each store, and for each region, etc).
- Generating surrogate-key values.
- Transposing or pivoting (turning multiple columns into multiple rows or vice versa).
- Splitting a column into multiple columns (For example, putting a comma-separated list specified as a string in one column as individual values in different columns).
- Applying any form of simple or complex data validation. If validation fails, it may result in a full, partial or no rejection of the data, and thus none, some or all the data is handed over to the next step, depending on the rule design and exception handling. Many of the above transformations may result in exceptions, for example, when a code translation parses an unknown code in the extracted data.

c) Load

The load phase loads the data into the end target, usually the Data Warehouse (DW). Depending on the requirements of the organization, this process varies widely. Some data warehouses may overwrite existing information with cumulative, updated data every week, while other DW (or even other parts of the same DW) may add new data in a histories form, for example, hourly. The timing and scope to replace or append are strategic design choices dependent on the time available and the business needs. More complex systems can maintain a history and audit trail of all changes to the data loaded in the DW. As the load phase interacts with a database, the constraints defined in the database schema — as well as in triggers activated upon data load — apply (for example, uniqueness, referential integrity, mandatory fields), which also contribute to the overall data quality performance of the ETL process.

II. ETL REQUIREMENTS

A practical and secure optimization of workflow in ETL which must satisfy the following basic requirements which can be explored as follows [1], [2], [3]:

ETL stands for extract, transform and load, the processes that enable companies to move data from multiple sources reformat and cleanse it, and load it into another database, a data mart or a data warehouse for analysis, or on another operational system to support a business process. Companies know they have valuable data lying around throughout their networks that needs to be moved from one place to another such as from one business application to another or to a data warehouse for analysis. The only problem is that the data lies in all sorts of heterogeneous systems, and therefore in all sorts of formats. For instance, a CRM (Customer Relationship Management) system may define a customer in one way, while a back-end accounting system may define the same customer differently. To solve the problem, companies use extract, transform and load (ETL) software, which includes reading data from its source, cleaning it up and formatting it uniformly, and then writing it to the target repository to be exploited. The data used in ETL processes can come from any source: a mainframe application, an ERP application, a CRM tool, a flat file, an Excel spreadsheet—even a message queue. Extraction can be done via Java Database Connectivity, Microsoft Corporation's Open Database Connectivity technology, proprietary code or by creating flat files. After extraction, the data is transformed, or modified, depending on the specific business logic involved so that it can be sent to the target repository. There are a variety of ways to perform the transformation, and the work involved varies. The data may require reformatting only, but most ETL operations also involve cleansing the data to remove duplicates and enforce consistency. Part of what the software does is, examines individual data fields and applies rules to consistently convert the contents to the form required by the target repository or application. In addition, the ETL process could involve standardizing name and address fields, verifying telephone numbers or expanding records with additional fields containing demographic information or data from other systems. The transformation occurs when the data from each source is mapped, cleansed and reconciled so it all can be tied together, with receivables tied to invoices and so on. After reconciliation, the data is transported and loaded into the data warehouse for analysis of things such as cycle times and total outstanding receivables. In the past, companies that were doing data warehousing projects often used homegrown code to support ETL processes. However, even those that had done successful implementations found that the source data file formats and the validation rules applying to the data evolved, requiring the ETL code to be modified and maintained. And companies encountered problems as they added systems and the amount of data in them grew. Lack of scalability has been a serious issue with homegrown ETL software. Providers of packaged ETL systems include Microsoft, which offers data transformation services bundled with its SQL Server database. Oracle has embedded some ETL capabilities in its database, and IBM offers a DB2 Information Integrator component for its warehouse offerings. More than half of all development work for data warehousing projects is typically dedicated to the design and implementation of ETL processes. Poorly

designed ETL processes are costly to maintain, change and update, so it is critical to make the right choices in terms of the right technology and tools that will be used for developing logic involved so that it can be sent to the target repository. The basic steps used for the development of the ETL Life cycle are as follows:

- 1. Cycle initiation.
- 2. Build reference data.
- 3. Extract (from sources).
- 4. Validate.
- 5. Transform (clean, apply business rules, check for data integrity, create aggregates).
 - 6. Stage (load into staging tables, if used).
- 7. Audit reports (for example, on compliance with business rules. Also, in case of failure, helps to diagnose/repair).
 - 8. Publish (to target tables).
 - 9. Archive.
 - 10. Clean up.
- 1) Architecture of Connector

Basic architecture of this connector will be:

The following operations will be performed by the connector;

- Reading data from source to the Quick ETL buffer.
- Writing data to target from the Quick ETL buffer.
- Managing the Meta data.

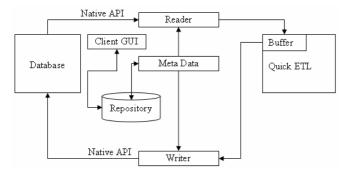


Figure 1 - Architecture of Connector

Various components of the connector are:

- · Reader.
- Writer
- Client GUI.

III. GENETIC ALGORITHMS

Melanie, M., in his book "An Introduction to Genetic Algorithms" stated that generally speaking, genetic algorithms are simulations of evolution, of what kind ever. In most cases, however, genetic algorithms are nothing else than probabilistic optimization methods which are based on the principles of evolution. He further suggested that if there is a solvable problem, a definition of an appropriate programming language, and a sufficiently large set of representative test examples (correct input-output pairs), a genetic algorithm is able to find a program which (approximately) solves the problem [12].

Goldberg, D in his book "Genetic Algorithms in Search, Optimization and Machine Learning" stated that crossover encourages information exchange among different individuals. It helps the propagation of useful genes in the population and assembling better individuals. In a lower level evolutionary algorithm, the crossover is implemented as a kind of cut-and-swap operator[6].

In 2003 Ulrich Bodenhofer suggested that in solving the problems related to genetic algorithms the following steps can be taken [4]:

Algorithm

t := 0;

Compute initial population B0;

WHILE stopping condition not fulfilled DO

BEGIN

select individuals for reproduction;

create off springs by crossing individuals;

eventually mutate some individuals;

compute new generation

END

As obvious from the above algorithm, the transition from one generation to the next consists of following basic components:

Selection: Mechanism for selecting individuals (strings) for reproduction according to their fitness (objective function value).

Crossover: Method of merging the genetic information of two individuals; if the coding is chosen properly, two good parents produces good children.

Mutation: In real evolution, the genetic material can by changed randomly by erroneous reproduction or other deformations of genes, for example, by gamma radiation. In genetic algorithms, mutation can be realized as a random deformation of the strings with a certain probability. The positive effect is preservation of genetic diversity and, as an effect, that local maxima can be avoided.

Following inferences were drawn about the Genetic Algorithms from the study and research carried by Bodenhofer:

- 1. Genetic Algorithms manipulate coded versions of the problem parameters instead of the parameters themselves, i.e. the search space is S instead of X itself.
- 2. Genetic Algorithms use probabilistic transition operators while conventional methods for continuous optimization apply deterministic transition operators. More specifically, the way a new generation is computed from the actual one has some random components
- 3. Normal genetic algorithms do not use any auxiliary information about the objective function value such as derivatives. Therefore, they can be applied to any kind of continuous or discrete optimization problem. The only thing to be done is to specify a meaningful decoding function.
- 4. While almost all conventional methods search from a single point, Genetic Algorithms always operate on a whole population of points (strings). This contributes much to the robustness of genetic algorithms. It improves the chance of reaching the global optimum and vice versa, reduces the risk of becoming trapped in a local stationary point.

Banzhaf [1999] had drawn the inference that Evolutionary algorithms have been shown to solve many real world problems. They use population based stochastic search strategies and are unlikely to be trapped in a poor local optimum. They make few assumptions about a problem domain yet are capable of incorporating domain knowledge in the design of chromosome representation and variation operators. They are particularly suited for large and complex problems where little prior knowledge is available [2].

The materialized view selection based on multiple query processing plans is a hard combinatorial optimization problem. Good selection of materialized views can only be found by taking a holistic approach and considering the optimization of both global processing plans and materialized view selection. A two-level structure for materialized view selection should be followed so as to get the proper result. It has facilitated greatly the development of several hybrid algorithms. In this literature survey the inference is drawn that there are several hybrid heuristic and evolutionary algorithms.

Pure evolutionary algorithms were found to be impractical due to their excessive computation time. Pure heuristic algorithms were unsatisfactory in terms of the quality of the solutions they found. Hybrid algorithms that combine the advantages of heuristic and evolutionary algorithms seem to perform the best. It show that applying an evolutionary algorithm to either global processing plan optimization or materialized view selection for a given global processing plan can reduce the total query and maintenance cost significantly. This is further revealed from the literature survey carried out that simply combining or merging optimal local processing

plans will not produce an optimal global processing plan in most cases. Finding an optimal global processing plan with optimal materialized views requires a two level hierarchy is needed. While the hybrid algorithms perform better than the heuristic algorithm in terms of cost savings, they often require longer computation time. While the heuristic algorithm typically took seconds to run, a hybrid algorithm typically took minutes, or even hours to run. Finding the suitable tradeoff between the computation time and the cost saving is the basic concept which we have to maximize.

Once a data warehousing design is completed and implemented, it will be used frequently and may last for a long time. Hence it is very important to optimize the design as much as possible, even if this means a relatively long design time. For optimizing the ETL process in the data ware houses we have to set up the theoretical framework for the problem, by modeling the problem as a state space search problem, with each state representing a particular design of the work flow as a graph. Since the problem is modeled as a state space search problem, it is the mandatory requirement to define transitions from one state to another.

IV. EFFICIENCY IMPROVEMENT IN READER WRITER PROBLEM

Job of the reader is to get data from the source system and give it to transformation unit for processing. In the transformation stage this data is processed and finally written to the target system.

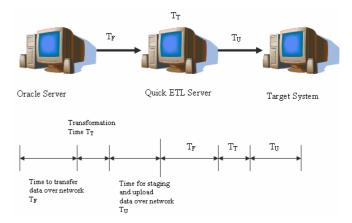


Figure 2 - ETL workflow execution timing.

Typical steps to read data from Source and transfer it to transformation stage consist of following steps:

- Fetch data from oracle server to local buffer.
- Transform the data type.
- Fill the target buffer.
- Flush the buffer.
- Repeat till data is available on oracle server.

Above mechanism to read and process data has following issues:

- When data is fetched from oracle server to local buffer, CPU is almost idle.
- During data transformation network bandwidth is not utilized.
 - During flushing of data CPU is idle.

The network bandwidth and CPU time can be efficiently used using two buffers instead of one buffer. Second buffer can be filled while first buffer is being processed and after the processing of first buffer, second buffer will be ready to process so waiting time will reduce. The steps will be:

- Fetch data into first buffer.
- While transformation and flushing fetch the data from oracle server to second buffer.
- Adjust the buffer size so that time to fetch is same as time for transformation and flushing.

Using two buffers following scenario exist:

- Fetch time is less than sum of transformation time and loading time.
- Fetch time is greater than sum of transformation time and loading time.
- Fetch time is same as sum of transformation time and loading time.

For the optimal utilization of network and CPU time, Fetch time should be same as sum of transformation time and loading time.

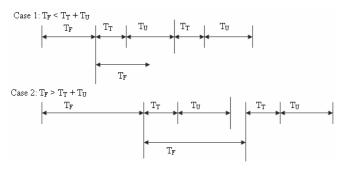


Figure 3 - Scenario for Two Buffer Implementation of Reader

To make the fetch time same as transformation time + loading time, the buffer size is changed dynamically. In the proposed algorithm two threads are used, one for fetching and other for transformation and loading. Both will work as:

Thread to fetch data:

- Check the status of buffer1.
- If empty fetch data in buffer1 and change the status to filled.

- Update the time TF1 used to fill buffer.
- If no more data available stop.
- Check the status of buffer2.
- If empty fetch data in buffer2 and change the status to filled.
 - Update the time TF2 used to fill buffer.
 - If no more data available stop.

Thread for transformation and uploading:

- Check the status of buffer1.
- If filled then transform data and flush it.
- Change the status to empty.
- Note the time for transformation and flushing, TP1 = TT1 + TU1.
 - If TF1 < TP1 increase the buffer size.
 - Check the status of buffer2.
 - If filled then transform data and flush it.
 - Change the status to empty.
- Note the time for transformation and flushing, TP2 = TT2 + TU2.
 - If TF2 > TP2 decrease the buffer size.

Adjusting buffer size:

- Buffer Size: S
- IF abs (TF TP) > 2 // Minimum Time Difference

o IF TF > TP

$$\partial S = S * TF / TP * Rnd$$
 // Random Value
between 0 & 1
New Buffer Size = S - ∂S

○ IF TF < TP

$$\partial$$
S= S * TP / TF * Rnd
New Buffer Size = S + ∂ S

V. WORKFLOW OPTIMIZATION

To implement the ETL workflow as state-space search problem the workflow is represented by a directed acyclic graph, where each node represents a transformation and edge represents the workflow. Consider the example: Two different databases contain the purchase information from different vendors. The database contains date, amount and vendor id. Now both the data need to be merged to single database so the possible workflow can be as in Figure 4. Where S is source node, SK is surrogate key assignment, U is union, SL is selection transformation and T is target.

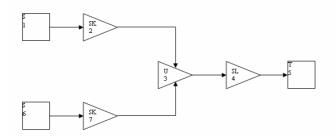


Figure 4- Initial Workflow

Solution Encoding

The solution for the current problem is encoded as an array of nodes. Each node represents a transformation. For example (1,2,6,7,3,4,5).

Initial Population

Initial population is generated by random transformation on the given workflow.

Fitness Function

The objective function of the individual is based on the total cost to execute the workflow. The cost of any transformation is defined by the cost model chosen. The problem is to minimize the cost, so an fitness function is derived which needs to be maximized.

Operators Selection

The tournament selection is used to generate new population.

Crossover

Single-point cross over is implemented to test the results. In the single-point crossover a random node is chosen and the parents are cut at crossover points and then join to produce the new individuals.

For example, consider following two individuals:

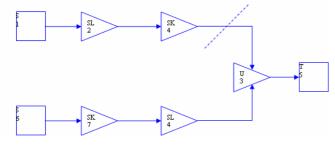


Figure 5 - Selected first individual

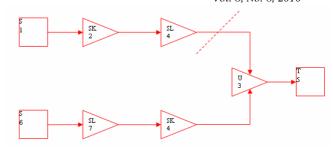


Figure 6 - Selected second individual

After crossover new individual will be

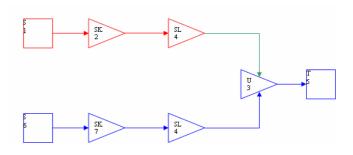


Figure 7- First individual after crossover

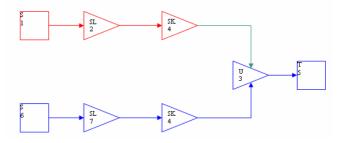


Figure 8 - Second individual after crossover

Mutation

Mutation with probability 1/3 is used. Mutation is applied as:

- Generate a random number.
- If random number is greater than mutation probability Exit.
- Find the list of transformations applicable to the node.
 - Select a random transformation.
 - Apply the transformation to the node.

In the experiment following transformation are supported:

- Swap.
- Factorize.

Distribute.

The algorithm used to generate the optimal workflow is as follows:

Initialize population from the given workflow by random mutation;

Do

Generate the Cost of each workflow

Make a log of lowest cost workflows

Determine which individuals should survive with fitness function;

Reproduce the survivors;

Select parents randomly from the survivors for crossover;

Select the crossover sites of the parents;

Produce the next new generation of workflows;

Mutate the new generation of workflows according to the mutation probability;

If iteration limit exceeded

output the optimal solutions

exit

endif

loop

A) Generating Initial Population

Initial population is generated from the given ETL workflow by randomly applying transitions on the workflow. The pseudo code for the initial population generation is:

While not pop_size

Copy the given workflow to create a new individual.

Select a random node.

Select a random transition.

Apply the transition to the selected node.

But during this processing network and CPU are not fully utilized. System can be made more efficient using more than one buffer as during processing time data can be fetched into another buffer. Workflow optimization is formulated as statespace search problem and state-space search is implemented using genetic algorithm.

B) Workflow Optimization

For the workflow following experiments have been conducted:

Experiment 1:

Initial Population 200

Crossover Simple

Selection Tournament Selection

Crossover Probability 0.1

Mutation Probability 0.9

Below is the average of 100 Runs for 10 generations.

Table 1: Generation No Vs Average Fitness

Generation No.	Average Fitness
0	396.9521
1	402.2411
2	408.1349
3	414.5278
4	420.2802
5	425.2696
6	430.4964
7	434.1143
8	440.2575
9	444.36096

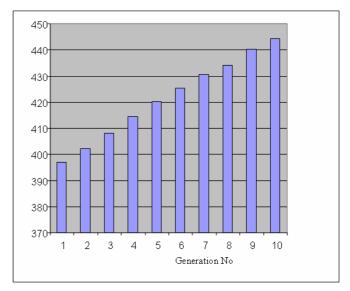


Figure 9 - Average Fitness Vs Generation No Experiment 2:

Initial Population 200

Crossover Simple

Selection Tournament Selection

Crossover Probability 0.9

Mutation Probability 0.1

Below is the average of 100 Runs for 10 generations.

Table 2: Generation No Vs Average Fitness

Generation No.	Average Fitness
0	397.9563
1	403.0553
2	409.2708
3	416.0089
4	420.0008
5	426.4026
6	431.6523
7	437.5481
8	443.3943
9	447.2889

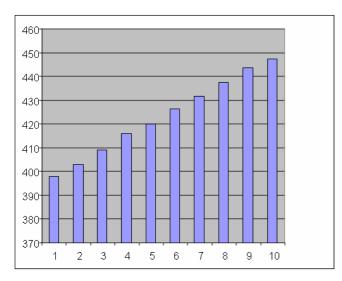


Figure 10 - Average Fitness Vs Generation No Comparative Study

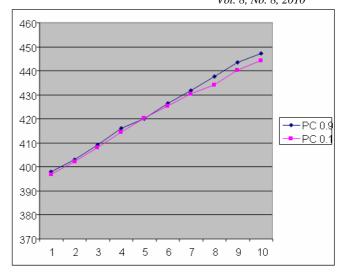


Figure 11- Average Fitness Vs Generation No for Different crossover probabilities

Experiment 3:

Random Initial Solution 200
Below is the average of 100 Runs for 10 generations.

Table 3: Generation No Vs Average Fitness

Generation No.	Average Fitness
0	397.9763
1	403.0553
2	397.9763
3	416.0089
4	397.9763
5	397.9763
6	431.6523
7	397.9763
8	416.0089
9	403.0553

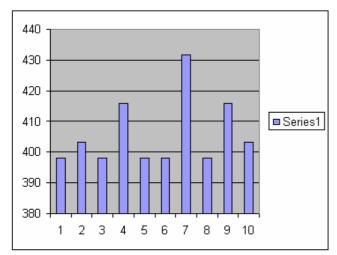


Figure 12 - Average Fitness Vs Generation No for Random Workflow generation

Comparison of GA Vs Random

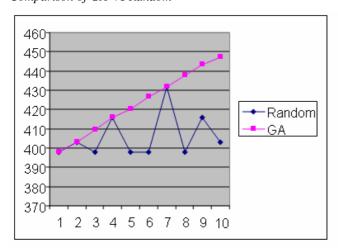


Figure 13: Comparison of Random workflow Vs GA Results:

This is clear from the above graph that as the generation goes on increases the average value of the fitness goes on increasing.

Limitations of The System

This approach is used to generate the optimal workflow as a search space solution. This technique does not consider the analytical analysis of the problem. Fitness is proportional to the total cost but does not include analysis of the relations between operators. From the results it is clear that genetic algorithm outperforms random workflow optimization. For the workflow optimization this study has used only swap, factorize and distribute transitions. It is further observed that as the generation in the genetic algorithm goes on increasing then accordingly the optimization of the ETL system increases. However the above conclusions only relate to the comparison of the random numbers and genetic algorithms.

There is a need to further investigate all the possible transitions, and their applicability conditions. It has not be derived that up to what stage the optimization will keep on increasing monotonically using the genetic algorithms or it will decrease after some time depending upon the complexity of the data warehouse. This can be the topic of further research that upto which stage the optimitality of the ETL system increases with the use of genetic algorithms.

VI. CONCLUSION

Good selection of materialized views can only be found by taking a holistic approach and considering the optimization of both global processing plans and materialized view selection. Pure evolutionary algorithms were found to be impractical due to their excessive computation time. Pure heuristic algorithms were unsatisfactory in terms of the quality of the solutions they found. Hybrid algorithms that combine the advantages of heuristic and evolutionary algorithms seem to perform the best.

For optimizing the ETL process in the data ware houses we have to set up the theoretical framework for the problem, by modeling the problem as a state space search problem, with each state representing a particular design of the work flow as a graph. Since the problem is modeled as a state space search problem, it is the mandatory requirement to define transitions from one state to another. For the workflow optimization this study has used only swap, factorize and distribute transitions. It is concluded that the results generated by optimizing the data of data warehouse using random numbers are less optimal as compared to when one uses the genetic algorithms. It is further observed that as the generation in the genetic algorithm goes on increasing then accordingly the optimization of the ETL system increases. However the above conclusions only relate to the comparison of the random numbers and genetic algorithms.

The efficiency of the reader is improved by the use of two buffers. The buffer size depends on the factors like network bandwidth, memory size, CPU speed etc. There is a need to further investigate the relation between these parameters and optimal number of buffers which can improve the performance. For the optimization of workflow the new workflows have been generated with the implementation of genetic algorithms. The cost of newly generated workflows produced by genetic algorithms is less than those produced by random way because the genetic algorithm explores the search space very fast with respect to the desired objective function.

VII. FUTURE SCOPE

There is a need to further investigate all the possible transitions, and their applicability conditions. It has not be derived that up to what stage the optimization will keep on increasing monotonically using the genetic algorithms or it will decrease after some time depending upon the complexity of the Data Warehouses (DW). This can be the topic of further research that up to which stage the optimitality of the ETL system increases with the use of genetic algorithms.

ACKNOWI EDGMENT

I (Raman Kumar) deeply indebted to my beloved master, supervisors, my parents and my research laboratory whose help, stimulating suggestions and encouragement helped me in all the time of research for and writing of this paper for journal. The authors also wish to thank many anonymous referees for their suggestions to improve this paper.

REFERENCES

- [1] Abadi, D., Carney, D. and Çetintemel, U., "A new model and architecture for data stream management", The VLDB Journal, 12(2) 2003, pp. 120-139.
- [2] Banzhaf, W., "Genetic and Evolutionary Computation Conference (GECCO)", Proceedings of the 1999 San Francisco, CA:Morgan Kaufmann, July 1999.
- [3] Baralis, E., Paraboschi, S. and Teniente, E., "Materialized view selection in a multidimensional database", in Proc. 23rd Int. Conf. Very Large Data Base (VLDB), 1997, pp. 156–165.
- [4] Bodenhofer, U., "Genetic Algorithms: theory and applications", Fuzzy logic lab linz, 2003.
- [5] Galhardas, H., Florescu, D., Shasha, D. and Simon, E., "Ajax: An Extensible Data Cleaning Tool", SIGMOD'00, Texas, 2000, pp. 590.
- [6] Goldberg, D., "Genetic Algorithms in Search, Optimization and Machine Learning", Reading, MA: Addison-Wesley, 1989.
- [7] Gupta, H., "Index selection for olap", in Proc. Int. Conf. Data Eng. (ICDE), 1997, pp. 208–219.
- [8] Gupta, H., Mumick, I., "Selection of views to materialize under a maintenance cost constraint", in Proc. Int. Conf. Database Theory (ICDT), 1999, pp. 453–470.
- [9] Ho, A., Lumpkin, G., "The genetic query optimizer", In Genetic Algorithms at Stanford 2004, J. R. Koza, Ed. Stanford, CA: Stanford Univ., 1994, pp. 67–76.
- [10] http://www.dbnet.ece.ntua.gr/~asimi/publications/SiVS04.pdf.
- [11] Ioannidis, Y., "Query optimization", ACM Comput. Surv., vol. 28, no.1, Mar. 1996, pp. 121–123.
- [12] Melanie, M., "An Introduction to Genetic Algorithms", MIT Press 1998.
- [13] Nicholas R. Jennings and Michael J. Wooldridge, "Agent Technology Foundations, Applications, and Markets", Springer-Verlag, 1998.
- [14] R.P. Majuca, W. Yurcik, and J.P. Kesan, "The Evolution of Cyber insurance", tech. report cs.CR/0601020, ACM Computing Research Repository, Jan. 2006.
- [15] Rahm, E. and Do, H., "Data Cleaning: Problems and Current Approaches", Bulletin of the Technical Committee on Data Engineering, 23(4), 2000.
- [16] Ross, K., Srivastava, D. and Sudarshan, S., "Materialized view maintenance and integrity constraint checking: Trading space for time", in Proc. ACM SIGMOD International Conference Manage Data, 1996, pp. 447–458.
- [17] S. Baer, "Rewarding IT Security in the Marketplace," Proc. 31st Research Conf. Comm., Information, and Internet Policy, TPRC, 2003; www.tprc.org/papers/2003/190/ BaerITSecurity.pdf.

- [18] Sellis, T., "Multiple-query optimization", ACM Trans. Database Syst., vol. 13, no. 1, Mar. 1999, pp. 23–52.
- [19] Simitsis, A., Vassiliadis, P and Sellis, T., "State-Space optimization of ETL Workflows", IEEE Transactions on Knowledge and Data Engineering, Vol 17, No 10, Oct 2005.
- [20] Simitsis, A., Vassiliadis, P. and Sellis, T., "Optimizing ETL Processes in Data Warehouse Environments (long version)", Available at
- [21] Stillger, M. and Spiliopoulou, M., "Genetic programming in database query optimization", in Proc First Annual Conference Genetic Programming, Stanford, CA, July 1996.
- [22] W. Baer and A. Parkinson, "Cyberinsurance in IT Management," IEEE Security & Privacy, vol. 5, no. 3, 2007, pp. 50–56.
- [23] Wisdom, J., "Research problems in data warehouse", in Proc 4th International Conference Inform. Knowledge Manage., 1995, pp. 25–30.
- [24] Zhang, C. and Yang, J., "Genetic algorithm for materialized view selection in data warehouse environments", in Proc. First Int. Conf. Data Warehousing Knowledge Discovery, Lecture Notes in Computer Science, Florence, Italy, 1999.
- [25] Zhang, C., Yao, X. and Yang, J., "An evolutionary approach to materialized views selection in a data warehouse environment", IEEE transactions on Systems, Man, and Cybernetics—Applications and Reviews, vol. 31, no. 3, August 2001, pp. 282-293.

AUTHOR'S PROFILE



Mr. Raman Kumar (er.ramankumar@aol.in) working as a Lecturer with the Department of Computer Science and Engineering, D A V Institute of Engineering and Technology, Jalandhar. Before joining D A V Institute of Engineering and Technology, Jalandhar,

He did his Bachelor of Technology with honours in Computer Science and Engineering from Guru Nanak Dev University; Amritsar (A 5 Star NAAC University). He did his Master of Technology with honours Computer Science and Engineering from Guru Nanak Dev University; Amritsar (A 5 Star NAAC University). His major area of research is Cryptography, Security Engineering and Information security. He has various publications in National as well as International Conferences and Journals on his research areas.

Vol. 8, No. 8, 2010

3D Protein Structure Comparison and Retrieval Methods: Investigation Study

Muhannad A. Abu-Hashem, Nur'Aini Abdul Rashid, Rosni Abdullah, Hesham A. Bahamish

School of Computer Science Universiti Sains Malaysia USM Penang, Malaysia

Abstract— The speed of the daily growth of computational biology databases opens the door for researchers in this field of study. Although much work have been done in this field, the results and performance are still imperfect due to insufficient review of the current methods. Here in this paper we discuss the common and most popular methods in the field of 3D protein structure comparison and retrieval. Also, we discuss the representation methods that have been used to support similarity process in order to get better results. The most important challenge related to the study of protein structure is to identify its function and chemical properties. At this point, the main factor in determining the chemical properties and the function of protein is the three dimensional structure of the protein. In other words, we cannot identify the function of a protein unless we represent it in its three dimensional structure. Hence, many methods were proposed for protein 3D structure representation, comparison, and retrieval. This paper summarizes the challenges, advantages and disadvantages of the current methods.

Keywords-3D protein structure; protein structure retrieval; protein structure comparison; PDB;

I. INTRODUCTION

Bioinformatics, considered a bridge connecting biology and computer science, is increasingly attracting the interest of researchers day by day. The size of protein, DNA and RNA databases is growing rapidly and as such necessitates the need for faster and efficient methods to manage and retrieve these data. expasy and rcsb [2, 3] are examples of protein databases websites which show the amount of the database growth every year. The goals of bioinformatics are to help biologists in collecting, managing, processing, storing, analyzing and retrieving genomic information that the biologists have and need [4]. One of the most interesting fields in bioinformatics is proteins where many researches focus on protein analyzing, predicting, comparison and similarity, retrieving, representation and more. The most important parts of proteins are its function and chemical properties which are determined at the 3D protein structure level [5]. So it is important to manage the data analyses, predict and retrieve the tertiary protein structure. Many researches have been carried out for this purpose. Furthermore, many databases (repositories) of protein structures are built to serve researchers in this field. One of the most common and essential protein structure repositories is the Protein Data Bank (PDB) [3, 6].

Most of the 3D protein structures in the database are determined using X-Ray crystallography methods and NMR [7]. These two methods are accurate but they are too slow and also too expensive. The first crystal 3D structure of protein myoglobin was determined and solved in 1958 [8].

Protein 3D structure similarity and retrieval importance increases day by day in tandem with the information that protein structures can provide and tell. Many methods have been proposed for protein structure representation, similarity and retrieval, but unfortunately the accuracy of the retrieval methods are still unsatisfactory. Those methods vary in techniques used in representation and comparison.

The benchmarks for the similarity and retrieval of the protein structures are the time of retrieving back a structure from the database and the accuracy of the retrieved structures. Most of the researches that have been done in solving this challenge consider DALI, VAST, computational extension (CE) and SCOP as a performance metrics.

1) PDB

Created and started in 1971 as an archive library for biological structures of macromolecules at Brookhaven National Laboratories [3, 6, 9]. The focus on PDB comes due to its importance and services in this domain where it is the most common database as well as it is considered as a primary database of protein 3D structures. The structures in PDB are obtained by two famous methods, X-Ray Crystallography and NMR [7], where they have been carefully validated. Figure 1 shows an example of the PDB database format.

understanding and identifying the functions of protein.

Vol. 8, No. 8, 2010

COMPND MOL_ID: 1; COMPND 2 MOLECULE: GLUTATHIONE SYNTHETASE; COMPND 3 CHAIN: A;

 ${\bf SOURCE~MOL_ID: 1;} \\ {\bf SOURCE~2~ORGANISM_SCIENTIFIC:~AVIAN~SARCOMA~VIRUS;} \\ {\bf TOTAL COMPART OF SARCOMA~VIRUS;} \\ {\bf SOURCE~2~ORGANISM_SCIENTIFIC:~AVIAN~SARCOMA~VIRUS;} \\ {\bf SOURCE~2~ORGANISM_SCIENTIFIC.~AVIAN~SARCOMA~VIRUS;} \\ {\bf SOURCE~2~ORGANISM_SCIENTIFIC.~AVIAN~SARCOMA~VIRUS~COMA~VI$

REMARK 3 REFINEMENT.
REMARK 3 PROGRAM : X-PLOR 3.851

Figure 1: PDB File Format Example [1]

Besides PDB database there are many databases that serve the 3D protein structures domain. The Structural Classification of Proteins, SCOP, is a protein structure database which describes the known evolutionary relationship of the protein structures as well as its structural relationship. It has been held at Cambridge University in the medical research council [10, 11]. The classification of protein Class, Architecture, Topology, and Homologous superfamily, CATH, [12] is a database that classifies the structures in the PDB hierarchically, and held at University of London. Furthermore, the Families of Structurally Similar Proteins (FSSP) database, built at the European Bioinformatics Institute, was created based on the DALI method [13, 14]. Moreover, a database called PROSITE [2, 15] is for the family classification of proteins where protein structures are classified into families that share the same functions.

2) Protein Structures

Proteins are the basic component of human cells as well as being the largest. So, the importance of proteins is clear regarding the role that proteins play in determining the function of cells. Proteins have many structures where each structure helps in the understanding the functions and chemical properties of living cells. The functions and chemical properties of proteins cannot be identified or determined before forming its tertiary structure. [16] shows the four levels of proteins starting from the amino acid sequences ending with its quaternary structure.

The trusted methods for identifying the tertiary structure of proteins are X-Ray Crystallography and NMR [7]. But the problems with those methods are cost and time where they are expensive and much time is massively consumed in order to form the tertiary structure.

3) Retrieval process

Searching for similar protein structures from the target database goes through many processes. First, the protein gets represented in a proper way that is suitable for comparison methods. This transformation of the protein has to be done for both the query protein structure and the database. This process is considered as a pre-process due to the size of the database and the time consumed by this stip. The rest of the subprocesses are all about how to get and measure the similarity and search for the query protein structure.

4) Problem Domain

Protein structure comparison and retrieval is one of the most important challenges in bioinformatics. Researchers' outputs in this field are still unsatisfactory where performance is less than the expected for time and accuracy. An advantage of protein structure retrieval is that it helps in predicting the tertiary structure of proteins and thus plays an important role in

The challenges in this domain are accuracy and time where faster and high accuracy methods are required without sacrificing the time. Many methods have been produced in this research area to find out the optimal solution for solving this challenge.

II. MATERIALS AND METHODS

A. Similarity Representaion Methods

Similarity representation of protein structure importance comes about due to its role in understanding the behavior of proteins. It helps in protein structure matching and similarity among other protein structures. Furthermore, it is the first step of protein structure comparison and retrieval. It is the process where the protein structure is built and rearranged in order to give simple and efficient representation for protein comparison to manage and efficiently prepare the matching. This data forming helps in fastening the comparison and retrieval process of proteins and has a high effect on the accuracy.

Many methods have been proposed for protein 3D structure similarity representation in order to enhance the comparisons of performance and efficiency. The following sections present these methods.

1) Matrix representation methods

This group uses matrices for presenting protein 3D structures. These methods are divided into two sub-groups, distance and similarity matrices.

a) Distance matrix: Two proteins are aligned in a matrix alike in order to represent them by calculating the distance between them. The values contained in the cells of the matrix represent the distance between the amino acids of the two proteins.

Holm L. and Sander C. [17] proposed an algorithm for protein structures comparison called DALI. The protein structures were represented as a distance matrix. The alignment between patterns and protein structures is done by executing a pairwise comparison on the distance matrices' patterns, where the similar patterns are kept in a list called pair list. Then, the patterns in the pair list are gathered to be aligned into a large set of pairs. The algorithm focuses on the subset of the patterns because of the size of the distance matrix, where it increases by increasing the length of the patterns or protein structures,. The distance matrix is reduced and the similar patterns are limited, in order to decrease the scope of the research process.

Aung Z. and Tan K.L [18] proposed a protein 3D structure retrieval system called PROTDEX2. The algorithm depends on index construction to represent the protein structure which is divided into two sub-processes, feature vectors extraction from

Vol. 8, No. 8, 2010

the contact regions (inter-SSE relationship) and constructing the inverted file index. The feature vectors are represented using distance matrix representation and SSEs (Secondary Structure Elements) vector representation. Each cell of the distance matrix contains the distance between the two $C\alpha$ atoms, where the distance is calculated using Euclidean distance. To calculate the SSE vectors' start and end points, equations adopted from [19] were used. Also STRIDE algorithm [20] has been used to identify the SSEs.

To construct the invert file index a 7-dimensional feature vectors and hash table were generated. Generating the 7-dimensional feature vectors is done first before these vectors are hashed into 7-dimensional hash table. Then, based on the generated hash table the inverted file index is built.

Masolo K. and Ramamohanarao K. [21] proposed a method for protein structure representation in order to accelerate the protein structure retrieval process. This method is based on constructing the protein feature vectors by using wavelet techniques. The idea behind using wavelet is its ability of compressing the application without sacrificing any of the details [22]. The earlier step of this method is building the distance matrix in order to build the feature vectors. The distance matrix is built by using the pairwise distance in the $C\alpha$ atoms level. To construct the representation of global structure they implemented the 2D decomposition of wavelet. Then the estimated coefficients are extracted from the upper part and the diagonal of the distance matrix.

b) Similarity matrix: [23] Similar to the distance matrix the two proteins are aligned in a matrix, but the values in the cells of the matrix will present similarity values between amino acids

Shindyalov I. N. and Bourne P.E [24] proposed an algorithm to enhance the protein structure's similarity and retrieval process. The first step in this algorithm is defining the alignment path. The alignment path can be defined as the longest continuous path in the similarity matrix by aligning two protein structures. Also, the algorithm took into consideration the alignment gaps, where it has conditions to control that which the two AFPs (Alignment Fragment Pairs) are aligned without gaps or one of the two proteins has gaps.

Chen S. C. and Chen T [25] proposed a protein structure retrieval method based on geometric hashing algorithm [26]. The pre-process for this algorithm is proteins feature extraction in order to get a new representation for the protein. For sequence alignment they adopted a similarity matrix called Dayhoff PAM250 [27] to enhance the performance of the algorithm.

2) Graph representation methods

Graph representation is one of the ways for protein 3D structure representation which is used to enhance the comparison and retrieval process for the protein structures.

Chen S. C. and Chen T [28] proposed a new algorithm for protein structure similarity and retrieval based on geometrical features. The algorithm represents the protein structure depending on the spatial relationship. It looks for the best alignment of the proteins first, and then it extracts the

geometric feature of the protein in order to define its geometric features.

Daras P. et al [29] proposed a three-dimensional shape structure comparison method for protein structure classification and retrieval. Protein structure representation in this algorithm is done by building a sphere and then triangulating it by using techniques of 3D modeling. By representing the protein as spheres, the number of connections and vertices will be reduced. Also in this step a new center of protein mass is calculated to be at the origin.

Sael L. et al. [30] introduced a novel algorithm for protein structure comparison and retrieval using 3D Zernike descriptors. Constructing the surface of the protein structures and detecting the surface area of the 3D structures are the initial steps of calculating the 3D Zernike descriptors. To build the protein surface, the algorithm first determines the surface area in the space of the structure. Furthermore, to calculate the Connolly surface (Triangle mesh) the algorithm depends on an existing program called MSROLL [31]. Then, the triangle mesh is arranged in the grid in a way that fits the protein in the grid.

B. Conventional Methods

The first step of protein structure retrieval from the Protein Data Bank (PDB) is protein structure comparison. If two proteins have the same structure, this implies that they might have the same function. Finding a protein similarity in the PDB helps the biologists to discover new functions for the proteins and also it helps in identifying unknown proteins functions. Many methods have been proposed in order to find the similarities between proteins. In this section we are going to present a preliminary study of the existing classified methods regarding their approaches.

1) Shape-Base approach

Sael L. et al. [30] introduced a novel algorithm for protein structure comparison and retrieval using 3D Zernike descriptors. 3D Zernike descriptors are used to help build the protein structure surface which provides a simpler representation of protein structures. Furthermore, this new representation helps to increase the speed of the comparison process. As a result of this research, searching for a protein in a database that consists of a few thousand protein structures takes less than a minute. The accuracy of the algorithm was 89.6% as compared with a well known algorithm in the domain called combinatorial extension (CE) [24].

3D Zernike descriptor focuses on the surface of the protein but not on the main chain which in some cases results in errors in the search results. This fault because of some structures has a similar surface shape but different main chain or similar main chain with different surface shape. Therefore, it is recommended that this algorithm is used as a primary filter preprocess for the protein structure comparison and retrieval methods.

Chen S. C. and Chen T. [25] proposed a protein structure retrieval method based on geometric hashing algorithm [26]. The idea of using the geometric hashing method is to find alike binding sites by applying surface matching on the protein. In

addition, a-hull and 3D reference frames techniques were adopted to simplify the algorithm complex computations.

As compared with [25], this algorithm obtained more accurate results. The experiments show that going through substructures for matching protein structures gives better results than matching between the whole protein as one block.

Daras P. et al. [32] proposed a three-dimensional shaped structure comparison method for protein structure classification and retrieval. Basically, the method depends on the geometry of the protein structures in the first stage, and in the second stage it depends on the primary and secondary structure of proteins. Next, when the 3D structure of the protein is built properly, the Spherical trace transform from [33] is used to build the geometry-based descriptor vectors. To evaluate the accuracy of the classification produced by using this method, FSSP/DALI database is used.

The accuracy of classification is calculated by the number of correctly predicted proteins over the real number of proteins in the database. The results of this method are compared with [34] to compare the accuracy of the classification and [25] to compare the retrieval performance by considering them close to the scope of the work. Regarding its accuracy, the proposed method outperformed the performance of [34] for a single domain chain. But in multiple domain proteins, the results were unconvincing. The performance of the proposed method shows some improvement.

Zhou Y. et al. [35] proposed a method that depends on using one of the shape distribution methods. It depends on the distribution of the backbone's $C\alpha$ representation. The method is proposed to measure the structural similarity between two proteins. It gives a strong shape distribution similarity among proteins with close structure and functions. The use of shape distribution is to calculate the distance between the atoms ($C\alpha$ atoms). After studying the shape functions they found that the most suitable shape function is D2 where it is easy to understand, faster, and robust.

2) Heuristic approach

Holm L. and Sander C. [17] proposed an algorithm for protein structures comparison called DALI. DALI algorithm is considered as a benchmark in the protein structure retrieval domain since many algorithms compare their results with it, in order to evaluate their results. The algorithm is looking for the best pairwise alignment [36-38] of proteins structures. To obtain the similarity among protein structures, the DALI algorithm looks for the occurrence of the query pattern in the protein structures database, then it identifies the largest common sequence among them. It gives high accurate and sensitive results regarding its search for similar structures of the two proteins, furthermore it does not get affected by the geometrical noise.

The DALI method shows high robustness and it is a general method which means it can be considered for solving other problems in the field. However, this method sacrifices the time in order to gain more accurate results with high sensitivity, where the searching process for one structure against the database takes almost a night.

Another well-known algorithm in this field is called CE (Combinatorial Extension) proposed by Shindyalov I. N. and Bourne P.E. [24] to enhance the protein structure similarity and retrieval process. CE algorithm is one of the best algorithms of protein retrieval regarding its accuracy of retrieving structures, but it saves time. Furthermore, the CE algorithm provides a good resolution of search in terms of finding protein structural similarity. Most of the methods in this field use CE algorithm to evaluate their results by comparing with CE results. The evaluation is in terms of accuracy and time of retrieving protein structures.

Aung Z. and Tan K.L. [18] proposed a protein 3D structure retrieval system called PROTDEX2. They came out with a system by adopting methods and techniques used successfully in information retrieval (IR) systems. The main objective behind this adaptation was to increase the speed of the searching process via protein database. The new system shows noticeable improvement in both sides, time and accuracy compared with PROTDEX [39] which is a previous version of PROTDEX2.

Speeding up the retrieving time is at the expense of accuracy. The results when compared with popular existing systems in this field which are DALI, CE, TOPSCAN [40] and the previous version of this work PROTDEX, PROTDEX2 produce the best retrieving time but not optimal results.

Chen S. C. and Chen T., [25] proposed an algorithm for protein structure similarity and retrieval based on geometrical features. Therefore, by identifying the geometrical features they avoid dealing with the chemical characteristics of the protein as well as the biological properties. To find the similarity among protein structures the algorithm focuses on getting the similarity of the appearance and spatial relationship of two sets of points. To speed up the process of comparison they identify the geometry features of the protein. Also, before extracting the features the algorithm searches for best alignment between the proteins. The new algorithm was compared with DALI and it was close in performance to DALI with more simplicity and efficiency.

III. CONCLUSION

In general, methods of 3D protein structure comparison and retrieval comes under one of two well known approaches which are heuristic and shape-base. Methods that follow the heuristic approach in general are fast. But on the other hand, they still have heuristic weaknesses as they sacrifice on accuracy. Shape-based methods focus on the robustness of the algorithm in order to reach higher accuracy and sensitivity with a reasonable running time. Anyhow, research in 3D protein structure comparison and retrieval field is still open while the current researches' results are still unconvincing. So, new innovative methods are needed to gain high performance in both accuracy and speed up.

ACKNOWLEDGMENT

This research is supported by the UNIVERSITI SAINS MALAYSIA (USM) and has been funded by "INSENTIF APEX".

REFERENCES

- [1] wwPDB, "Protein Data Bank Contents Guide: Atomic Coordinate Entry Format Description Version 3.20," 2008. " http://mmcif.pdb.org/dictionaries/mmcif_pdbx.dic/Index/index.html "
- [2] expasy, "Database of protein domains, families and functional sites," 2010. "http://www.expasy.ch/prosite/"
- [3] RCSB, "RCSB Protein Data Bank," 2010. " http://www.rcsb.org/pdb/"
- [4] C. Jacques, "Bioinformatics\— an introduction for computer scientists," ACM Comput. Surv., vol. 36, pp. 122-158, 2004.
- [5] C.-I. Branden and J. Tooze, *Introduction to Protein Structure*: Garland Publishing, 1999.
- [6] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," *Nucl. Acids Res.*, vol. 28, pp. 235-242, January 1, 2000 2000.
- [7] J. L. Sussman, J. J. D. Ling, N. O. Manning, J. Prilusky, O. Ritter, and E. E. Abola, "Acta Crystallogr," vol. 54, pp. 1078-1084, 1998.
- [8] M. M. Bluhm, G. Bodo, H. M. Dintzis, and J. C. Kendrew, "The Crystal Structure of Myoglobin. IV. A Fourier Projection of Sperm-Whale Myoglobin by the Method of Isomorphous Replacement," *Proceedings* of the Royal Society of London. Series A, Mathematical and Physical Sciences, vol. 246, pp. 369-389, 1958.
- [9] F. C. Bernstein, T. F. Koetzle, G. J. Williams, E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, "The Protein Data Bank. A computer-based archival file for macromolecular structures," *European journal of biochemistry / FEBS*, vol. 80, pp. 319-24, 1977.
- [10] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, "SCOP: A structural classification of proteins database for the investigation of sequences and structures," *Journal of Molecular Biology*, vol. 247, pp. 536-540, 1995.
- [11] L. Lo Conte, S. E. Brenner, T. J. P. Hubbard, C. Chothia, and A. G. Murzin, "SCOP database in 2002: refinements accommodate structural genomics," *Nucl. Acids Res.*, vol. 30, pp. 264-267, January 1, 2002 2002.
- [12] C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton, "CATH a hierarchic classification of protein domain structures," *Structure (London, England : 1993)*, vol. 5, pp. 1093-1109, 08/15 1997.
- [13] L. Holm and C. Sander, "The FSSP database: fold classification based on structure-structure alignment of proteins," *Nucl. Acids Res.*, vol. 24, pp. 206-209, January 1, 1996 1996.
- [14] EMBL-EBI, "The European Bionformatics Institute," 2009. " http://www.ebi.ac.uk/"
- [15] L. Falquet, M. Pagni, P. Bucher, N. Hulo, C. J. A. Sigrist, K. Hofmann, and A. Bairoch, "The PROSITE database, its status in 2002," *Nucl. Acids Res.*, vol. 30, pp. 235-238, January 1, 2002 2002.
- [16] genome, "National Human Genome Research Institute." vol. 2010, 2009. "http://www.genome.gov/Glossary/"
- [17] L. Holm, "Protein Structure Comparison by Alignment of Distance Matrices," *Journal of Molecular Biology*, vol. 233, pp. 123-138, 1993.
- [18] Z. Aung and K.-L. Tan, "Rapid 3D protein structure database searching using information retrieval techniques," *Bioinformatics*, vol. 20, pp. 1045-1052, May 1, 2004 2004.
- [19] P. S. Amit and L. B. Douglas, "Hierarchical Protein Structure Superposition Using Both Secondary Structure and Atomic Representations," in *Proceedings of the 5th International Conference on Intelligent Systems for Molecular Biology*: AAAI Press, 1997.

- [20] D. Frishman and P. Argos, "Knowledge-based protein secondary structure assignment," *Proteins: Structure, Function, and Genetics*, vol. 23, pp. 566-579, 1995.
- [21] M. Keith, P. Srinivasan, and R. Kotagiri, "Structure-based querying of proteins using wavelets," in *Proceedings of the 15th ACM international* conference on Information and knowledge management Arlington, Virginia, USA: ACM, 2006.
- [22] S. Mallat, A Wavelet Tour of Signal Processing, Second Edition (Wavelet Analysis & Its Applications): Academic Press, 1999.
- [23] D. J. States, W. Gish, and S. F. Altschul, "Improved sensitivity of nucleic acid database searches using application-specific scoring matrices," *Methods*, vol. 3, pp. 66-70, 1991.
- [24] I. N. Shindyalov and P. E. Bourne, "Protein structure alignment by incremental combinatorial extension (CE) of the optimal path," *Protein Eng*, vol. 11, pp. 739-747, 1998.
- [25] s. c. chen and t. chen, "protein retrieval by matching 3D surfaces," 2002.
- [26] Y. Lamdan and H. J. Wolfson, "Geometric Hashing: A General And Efficient Model-based Recognition Scheme," in *Computer Vision.*, Second International Conference on, 1988, pp. 238-249.
- [27] R. M. Schwartz and M. O. Dayhoff, "Atlas of Protein Sequence and Structure," *National biomedical research foundation Washington DC*, vol. 5, 353-358, 1978.
- [28] C. Shann-Ching and C. Tsuhan, "Retrieval of 3D protein structures," in Image Processing. 2002. Proceedings. 2002 International Conference on, 2002, pp. 933-936 vol.3.
- [29] D. Petros, Z. Dimitrios, A. Apostolos, T. Dimitrios, and S. Michael Gerassimos, "Three-Dimensional Shape-Structure Comparison Method for Protein Classification," *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, vol. 3, pp. 193-207, 2006.
- [30] L. Sael, B. Li, D. La, Y. Fang, K. Ramani, R. Rustamov, and D. Kihara, "Fast protein tertiary structure retrieval based on global surface shape similarity," *Proteins*, 2008.
- [31] M. L. Connolly, "Solvent-accessible surfaces of proteins and nucleic acids," *Science*, vol. 221, pp. 709-713, 1983.
- [32] P. Daras, D. Zarpalas, A. Axenopoulos, D. Tzovaras, and M. G. Strintzis, "Three-dimensional shape-structure comparison method for protein classification," *Ieee-Acm Transactions on Computational Biology and Bioinformatics*, vol. 3, pp. 193-207, Jul-Sep 2006.
- [33] D. Zarpalas, P. Daras, A. Axenopoulos, D. Tzovaras, and M. G. Strintzis, "3D model search and retrieval using the spherical trace transform," *Eurasip Journal on Advances in Signal Processing*, pp. -, 2007.
- [34] M. Ankerst, G. Kastenm, H.-P. Kriegel, and T. Seidl, "Nearest Neighbor Classification in 3D Protein Databases," in *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*: AAAI Press, 1999.
- [35] Z. Ying, Z. Kaixing, and M. Yuankui, "3D protein structure similarity comparison using a shape distribution method," in *Information Technology and Applications in Biomedicine*, 2008. ITAB 2008. International Conference on, 2008, pp. 233-236.
- [36] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology*, vol. 48, pp. 443-453, 1970.
- [37] W. J. Wilbur and D. J. Lipman, "Rapid Similarity Searches in Nucleic Acid and Protein Databanks. Proc. Natl. Acad. Sci. USA, Vol(80), 726-730, "1983
- [38] s. f. Altschul, w. Gish, w. Miller, w. e. Myers, and d. j. Lipman, "Basic Local Alignment Search Tool. J.Mol.Biol 215, 403-410," 1990.
- [39] Z. Y. Aung, W. Fu, and K. L. Tan, "An efficient index-based protein structure database searching method," *Eighth International Conference* on Database Systems for Advanced Applications, Proceedings, pp. 311-318, 2003.
- [40] A. C. R. Martin, "The ups and downs of protein topology; rapid comparison of protein structure," *Protein Engineering*, vol. 13, pp. 829-837, Dec 2000.

The Impact of Speed on the Performance of Dynamic Source Routing in Mobile Ad-Hoc Networks

Naseer Ali Husieen, Osman B Ghazali, Suhaidi Hassan, Mohammed M. Kadhum Internetworks Research Group College of Arts and Sciences University Utara Malaysia 06010 UUM Sintok, Malaysia

Abstract— Ad-hoc networks are characterized by multihop wireless connectivity, frequently changing network topology and the need for efficient dynamic routing protocols plays an important role. Due to mobility in Ad-hoc network, the topology of the network may change rapidly. The mobility models represent the moving behavior of each mobile node in the MANET that should be realistic. This paper concerns performance of mobile Ad-hoc network (MANET) routing protocol with respect to the effects of mobility model on the performance of DSR protocol for the purpose of finding the optimal settings of node speed. In this paper, we evaluate the performance of DSR protocol using Random Waypoint Mobility Model in terms of node speed, number of connections, and number of nodes.

Keywords-MANET, Mobility Models, Routing Protocol, DSR Protocol.

I. INTRODUCTION

With existing advances in technology, wireless networks are growing in popularity. Wireless networks permit users the freedom to move from one position to another without break of their computing services. Adhoc networks is one of the subset of wireless network that dynamically forming a temporary network without using any existing network infrastructure or centralized administration. A major problem in ad hoc network is how to send data packets among mobile nodes efficiently without fixed topology or centralized control, which is the most important goal of ad hoc routing protocols. Therefore, it is necessary a high-quality routing protocol in order to establish the link between the nodes, since the mobile node can vary their topology regularly. In ad-hoc network the routing protocol is one of the important issue and most challenging research area, since mobile ad-hoc network vary their topology frequently. Generally, the major task of routing in a network is to detect and keep the best path to send data packets between source and target through intermediate nodes. There are two categories of routing protocols in ad hoc networks: Protocols: In this type of protocols such as DSDV, OLSR, consistent and up to-date routing information to all nodes is maintain at each node. ii.

Reactive Protocols: In this type of protocols such as DSR, AODV, the routes are created when it's required to send data packets from the source to the destination [1]. We have determined the impact of four factors on the performance of DSR by using Random waypoint mobility model in our previous paper [2] in press. These factors pause time, network size, number of traffic sources and routing protocol. We examine the impact of these factors on four performance metrics: packet delivery ratio, average end-to-end delay, normalized routing load and protocol overhead. In this paper, we use Random waypoint as mobility model on DSR protocol to study the effect of node speed with other factors in order to find the optimal setting for the node speed parameter with different scenarios. For this performance study, we use Network Simulator 2 (ns-2) version 2.34.

II. DYNAMIC SOURCE ROUTING (DSR)

DSR is reactive and efficient protocol. It determines the correct path only when a packet wants to be forwarded. The node broadcast the network with a route request and builds the essential path from the responses it receives. DSR allows the network to be fully self configuring with no need for any existing network infrastructure or administration. The DSR protocol is composed of two main mechanisms that work together to allow the discovery and maintenance of source routes in the ad-hoc network. All aspects of protocol operate entirely on demand allowing routing packet overhead of DSR to scale up automatically [3] [4].

Route Discovery: The example for route discovery shown in Figure 1. When a source node 1 wants to send data packets to the destination node 8, node 1 will broadcast Route Request Packet (RREQ) to all the neighbor nodes 2, 3, 4. After intermediate nodes receive theses packets will rebroadcast these packets to the destination if there is no route in the route cache. When the destination node 8 will receive RREQ, node 8 will inform the source node 1 by sending the Route Reply Packet (RREP). The source node will start sending the data packets to the destination through the intermediate nodes. This process mechanism called route discovery.

Route Maintenance: This mechanism contains two packets; Route Error Packet (RERR) and ACKs packets. Route error packet generated when there is changing in the network or node out of the transmission range which causes link failure. Intermediate node will send RERR to the source node. Source will check if there is route in the route cache to send the packets to the destination, if there is no alternative route. Source node will reinitiate route discovery process again. These processes will take long delay to re-establish again in order to send the data packets to the destination [5].

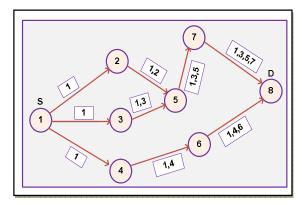


Fig.1. DSR Route Discovery

III. RANDOM WAY POITN MODEL

The Random waypoint model is widely and simply used to evaluate the performance of ad hoc routing protocols. The implementation of this model in the network simulator (ns-2) is as follows: each mobile node arbitrarily selects one position in the simulation field as the target, then moves towards this target with fix velocity selected uniformly and randomly from [0, Vmax], where the parameter Vmax is the maximum velocity for each mobile node [6]. The velocity and path of the nodes are selected separately from of other nodes. When will reaching the target, the node stops for a period of time defined by the 'pause time'.

In the Random waypoint model, velocity and pause time are two key parameters that determine the mobility performance of nodes. When the pause time is long and velocity is small the topology of ad-hoc network becomes stable. On the other hand, when the mobile node moves fast and the pause time is small; the topology is likely to be highly dynamic. Random waypoint model can create different mobility scenarios with different levels of node speed. In Figure 2, shows that node movement in the Random waypoint.

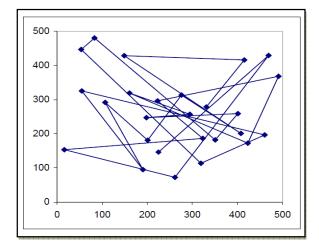


Fig. 2. Traveling pattern of an MN using the Random Waypoint Mobility Model

IV. RELATED WORK

In the recent years, several works has been done by Perkins, Hughes and Owen [7] shows that some parameters such as pause time, node speed, increasing the number of nodes, and increase the number of sources can have an effect on the routing protocols performance. In their work Random waypoint model has been used, but employed Global Mobile System Simulator (GloMoSim) rather than *ns-2*.

Azzedine Baoukerche has comparing four routing protocols such as (AODV, PAODV, DSR, and CBRP) [8]. The simulation parameters were tested in his paper with maximum number of nodes 25 and low traffic with maximum speed 20 m/s. However, our work tested with various numbers of nodes (10, 20, 40, and 80) with different source connections (4, 8, 30, and 40) which can make high traffic and various speed 20, 40, 60, and 80.

Yogesh, Yudhvir, and Manish, they have compared and analysis two reactive routing protocols such as AODV, DSR [9]. The main objective in their paper to evaluate the performance of these two protocols based on the packet delivery fraction, end –to-end delay, and normalized routing load. The simulation parameters were tested increasing number on nodes and various pause times with fixed maximum speed (0-25 m/s only). While our work with various maximum speeds in order to select the optimal setting for maximum speed.

V. SIMULATION SETUP

The MANET network simulations are implemented using Random waypoint model which can generate by using movement tool (setdest) in ns-2 simulator. The simulation period for each scenario is 200 seconds and the simulated mobility network area is 1000 m x 500 m rectangle. Simulation runs are made with the number of random traffic Constant Bit Rate (CBR) which can generate by using (cbgen.tcl). Figure 2, it shows that the

simulation methodology for our implementation. The rest of simulation parameters shown in the Table below.

TADIEI	CIMILIA	ATION PAR	AMETEDS

Parameters	Value
Simulation Time	200 s
No. of Nodes	10, 20, 40, 80
No. of connections	4, 8, 30, 40
Pause Time	40 s
Simulation Area	1000 x 500 m
Traffic Type	Constant Bit Rate (CBR)
Maximum Speed	20 ,40,60,80 m/s
Mobility Model	Random Waypoint
Routing Protocol	DSR
MAC Type	802.11

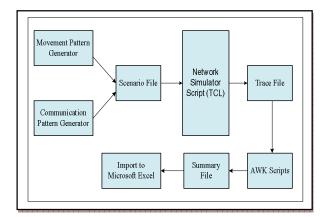
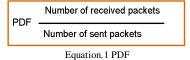


Fig.3. Simulation Methodology

VI. PERFORMANCE METRICS

We have consider packet delivery ratio, end to end delay, protocol control overhead and normalized routing load as a metrics during our simulation in order to evaluate the performance of the DSR protocol.

Packet Delivery Fractions (PDF): the packet delivery ratio is calculated by dividing the number of packets received by the destination through the number of originates packets by the application layer of the initiator. PDF is specifying the loss packets rate, which limits the maximum throughput over the entire network. The more complete and correct routing protocol, the better packet delivery ratio.



Average end to end delay: this is defined as the average delay in transmission of a packet between two nodes and is calculated as follows:

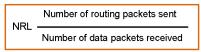
$$AED = \sum_{i=0}^{n} \frac{\text{Time packet received}_{i} - \text{time packet sent}_{i}}{\text{Total number of packets received}}$$

Equation.2 AED

A higher value of end-to-end delay means that the network is congested and hence the routing protocol does not perform well. The upper bound on the values of end-to-end delay is determined by the application.

Protocol Control Overhead: This is the ratio of the number of protocol control packets transmitted to the number of data packets received.

Normalized routing load: this is calculated as the ratio between the numbers of routing packets transmitted to the number of packets actually received (thus accounting for any dropped packets):



Equation.3 NRL

This metric gives an analysis of routing protocol efficiency, since the number of routing packets sent per data packet gives an idea of how well the protocol maintains the routing information updated. The lower of NRL, the lower the overhead of routing protocol and consequently the higher the efficiency of the protocol.

VII. RESULTS AND DISCUSSION

In this section, details of the simulation results in term of packet-delivery fraction, average end to end delay, protocol overhead, and normalized routing load. All the results were obtained by averaging 5 times over the simulation for every scenario in order to select the optimal setting for the maximum speed.

i. Packet Delivery Fraction (PDF)

This metric with high packet delivery ratio, routing protocol will be more efficient. Figure 4, shows that packed delivery ratio is decreased in first scenario whenever increasing node speed. As above stated, it has been taken four main scenarios and each of this main contains four sub-scenarios which means 4x4 = 16 scenarios have been taken in this experiment. In each of these experiments of sub-scenarios, the node speed increases while other parameters are constant. In Figure 4, shows that the optimal setting is 20 speeds among all of the four scenarios in term of packet delivery ratio.

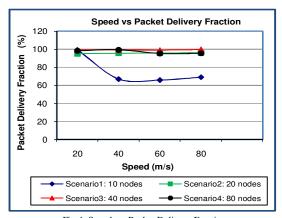


Fig.4. Speed vs. Packet Delivery Fraction

ii. Average End-to-End Delay

The average end-to-end delay is affected when the traffic Constant Bit Rate (CBR) is high rate of packets as well. The buffers become filled much faster, so the packets have to wait in the buffers a much longer period of time before they are sent. Figure 5, shows that effect of node speed on the end to end delay. In the first scenario 10 nodes with 4 connections CBR, there is less delay with 20 speeds comparing with others sub -scenario 40, 60, and 80 speeds. In addition when the number of mobile nodes (MNs) is increased to 80 nodes and 40 CBR connections with respect increase node speed up to 80 m/s the end to end delay increases because of the time consumed for route discovery and the increasing number of packets in the buffer. However, when the pause time is 40 s and speed 20 m/s, the network is stable and the end to end delay decreases. With normal speed, the end to end delay is low because the network is not congested.

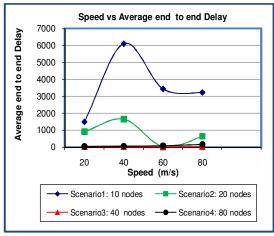


Fig.5.Speed vs. Average end to end delay

iii. Routing Protocol Overhead

In Figure 6, it shows the overhead result which generated by the routing protocols to achieve this level of data packet delivery. Figure 6, shows that overhead is direct proportional to the number of sending packets, in the first scenario with low mobility overhead increased whenever node speed increased. With normal speed 20 m/s in first scenario overhead decreased. The rest of scenarios with high mobility overhead is expect to increase and decrease because there are more destinations to which the network must maintain working routes.

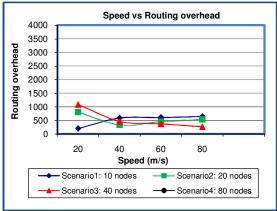
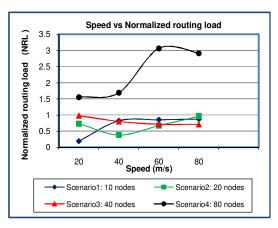


Fig.6.Speed vs. Routing overhead

iv. Normalized Routing Load (NRL)

In figure 6, the value of normalized routing load versus node speeds are plotted. From the figure 7, this is clear that DSR protocol performs well in the first scenario with node speed 20 m/s. Because of the NRL direct proportional with overhead and sending packets.NRL represents the number of routing packets transmitted per data packet delivered at the destination. This metrics checks the efficiency of the DSR protocols, meaning that with low NRL, DSR perform well.



 $Fig.7.Speed\ vs.\ Normalized\ routing\ load$

VIII. CONCLUTION AND FUTURE WORK

The most important parameter of a mobility model is a node speed, either in the form of a constant value or in the form of a certain distribution. In this paper, we have present our work on evaluating performance of DSR protocol under widely and simplest mobility model which called Random Waypoint with respect to the effect of node speed for different scenarios. Simulation result shows that node speed has significant on the performance of DSR protocol because mobility models are characterized by the movement of their constituents. The nature of movement its speed, direction, and rate of change can have a dramatic effect on protocols and systems designed to support mobility. The Mobile nodes (MNs) randomly select the next destination in the simulation area and choose a speed uniformly distributed between the minimum and maximum speed and travels with a random speed which is chosen uniformly. In addition, results shows that average optimal setting for our scenarios is when node speed is 20 m/s. The experimentation suggests that several parameters such as maximum speed, node density pause time and traffic source of nodes also affect the routing performance and need to be investigated with various mobility models.

REFERENCE

- [1] Elizabeth Belding -Royer," Routing approaches inn mobile ad hoc networks", in: S.Basagni, M.Conti, S.Giordano, I.Stojemenvoic (Eds), Ad Hoc Networking, IEEE Press Wiley, New York, 2003.
- [2] Naseer,Osman, Suhaidi, Kadhum "Effect of Pause Time on the Performance of Mobile Ad Hoc Network Routing Protocols". in IEEE 4International Conference on Inteligent Information Technology Application (IITA 2010), China, 2010, in press
- [3] D.B. Johnson, D.A. Maltz, Y Hu, Dynamic Source Routing Protocol for Mobile Ad-hoc Networks (DSR), http://www.ietf.org/internet-drafts/draft-ietf-manet-dsr-10.txt, July 2004.
- [4] David B. Johnson David A. Maltz Josh Brooch, "DSR: the Dynamic Source Routing Protocol for Multi-Hop Wireless Ad Hoc Networks", http://www.monarch.cs.cmu.edu/.
- [5] The Dynamic Source Routing Protocol for Mobile Ad-hoc Networks, "http://www.ietf.org/internet-drafts/draft-ietf-manetdsr-03.txt, IETF Internet draft, Oct. 1999.
- [6] L. Breslau, D. Estrin, K. Fall, S. Floyd, J. Heidemann, A. Helmy, P. Huang, S. McCanne, K. Varadhan, Y. Xu, and H. Yu, Advances in network simulation, in IEEE Computer, vol. 33, no. 5, May 2000, pp. 59--67.
- [7] D. Perkins, H. D. Hughes, and C. B. Owen, "Factors affecting the performance of ad-hoc networks," in Proceedings of the IEEE International Conference on Communications (ICC), Electronic Publication: Digital Object Identifiers (DOIs), 2000.
- [8] Azzedine Boukerche," Performance Evaluation of Routing Protocols for Ad Hoc Wireless Networks", Kluwer Academic Publishers. Manufactured in the Netherlands Mobile Networks and Applications 9, 333–342, 2004.
- [9] Yogesh, Yudhvir, and Manish," Simulation based Performance Analysis of On-Demand Routing Protocols in MANETs, Second international Conference on Computer Modeling and Simulation.

AUTHORS PROFILE



Ali Naseer Husieen, received his B.Sc. degree in Computer Science from Al-Rafedain University, Iraq and, his M.Sc. degree in Computer Science focusing on Computer Network and Communications from Hamdard University, Delhi (Faculty of Computer Science), India. Naseer

currently attached to the InterNetWorks Research Group, College of Arts and Sciences at the Utara University Malaysia. He is currently pursuing his PhD research in Ad-hoc Mobile networking as a doctoral researcher. His current research interest is on Ad-hoc mobile network routing protocol.



published a number conferences.

Osman Ghazali, Ph.D. a
Senior Lecturer in the
Department of Computer
Science for Postgraduate
Studies, Northern University
of Malaysia (Universiti
Utara Malaysia). He
received his BIT, Master
and PhD in Information
Technology from Northern
University of Malaysia in
1994, 1996, and 2008. He

of papers in international



Associate Professor Suhaidi Hassan is currently the Assistant Vice Chancellor of the College of Arts and Sciences, Universiti Utara Malaysia (UUM). He is an associate professor in Computer Systems and Communication Networks and the former Dean of the Faculty Information

Technology, Universiti Utara Malaysia. Dr. Suhaidi Hassan received his B.Sc. degree in Computer Science from Binghamton University, New York (USA) and his MS degree in Information Science (concentration in Telecommunications and Networks) from the University of Pittsburgh, Pennsylvania (USA). He received his PhD

degree in computing (focusing in Networks Performance Engineering) from the University of Leeds in the United Kingdom. In 2006, he established the ITU-UUM Asia Pacific Centre of Excellence (ASP CoE) for Rural ICT Development, a human resource development initiative of the Geneva-based International Telecommunication Union (ITU) which serves as the focal point for all rural ICT development initiatives across Asia Pacific region by providing executive training programs, knowledge repositories, R &D and consultancy activities. Dr. Suhaidi Hassan is a senior member of the Institute of Electrical and Electronic Engineers (IEEE) in which he actively involved in both the IEEE Communications and IEEE Computer societies. He has served as the Vice Chair (2003-2007) of the IEEE Malaysia Computer Society. He also serves as a technical committee for the Malavsian Research and Educational Network (MYREN) and as a Council Member of the Cisco Malaysia Network Academy.



Mohammed M. Kadhum, Ph.D. is an assistant professor in the Graduate Department of Computer Science, Universiti Utara Malaysia (UUM) and is currently attached to the InterNetWorks Research Group at the UUM College of Arts and Sciences as a research advisor. He had completed his PhD research

in computer networking at Universiti Utara Malaysia (UUM). His research interest is on Internet Congestion and QoS. He has been awarded with several medals for his outstanding research projects. His professional activity includes being positioned as Technical Program Chair for NetApps2008 and NetApps2010, a technical committee member for various well known journal and international conferences, a speaker for conferences, and a member of several science and technology societies. To date, he has published a number of papers including on well-known and influential international journals.

Multidimensionality in Agile Software Development

Ashima, Assistant Professor, Computer Science and Engineering Department
Thapar University, Patiala

Dr. Himanshu Aggarwal, Associate Professor. Faculty of Computer Engineering,
Punjabi University, Patiala

Abstract: Among new software development processes, Agile Software Development (ASD) gives the software industry a new idea of quick and timely delivery of product. Agile methodologies got overwhelming response by all levels of software organizations. But limited scope of software designing and reusability of components do not let it to be made first choice of software development and professionals. Agility Multidimensional constraints like software design and Reusability, architecture and risk, iterations and changeability. Rapid development combined changeability at later phases adds charm to ASD but missing designing and reusability act as a hurdle. Popularity of any software product is actually in length of its stay in market that ofcouse yields them rewards in terms of money compared to their investments. Agility's approach of development towards *specialized* components also lessens their probability of staying long in market. This paper aims to find how reusability by adding a bit of designing and developing specialized cum generalized components can be achieved in ASD.

Introduction: Agile Software Development methods and techniques are being followed in the industry from the last decade to get quality product and to reduce development time. Rapid development and accommodate changes at any level of development gives the competitive advantage to the Agile processes over Traditional processes. But to get best the combination of both the processes is required. A proper degree of specialization and generalization needed to be maintained. Inclusion of architecture specific designing in ASD can make it a reliability prone approach i.e. ASD without risk.

Reusability also contributes towards quality product and the rapid development. [19] reveals that Japanese projects also exhibited higher levels of reuse while spending more time on product design as compared to American teams which spend more time on actual coding and concludes that Indian firms are doing great job in combining conventional best practices, such as specification and review, with more flexible techniques that should enable them to respond more effectively to customer demands. If such a trend is replicated across the broader population, it suggests the Indian software industry is likely to experience continued growth and success in future.

Progression to Agile Software Development

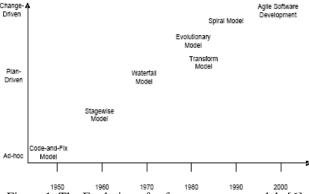


Figure 1: The Evolution of software process models [6]

Figure 1 shows the evolution of software development processes. Waterfall model was being followed where requirements are fixed and the next phase starts when the earlier one finished. It's representative of the traditional methods. To overcome the limitations of waterfall model, evolutionary model and spiral model comes into picture where prototype is first made and then that is converted to the working software. But all have one common limitation that no process could handle the change of requirements at later phases. Agile development which include many methodologies as XP, SCRUM, Lean Software Development, FDD,DSDM is being accepted in industry because of adaptation to change even at the later stages of the development and also for rapid development.

Any method to be agile the values and principles of the Agile Manifesto (Agile Alliance 2001) set out the central elements of agility. "We are uncovering better ways of developing software by doing it and helping others do it. Through this work we have come to values: Individuals and interactions over processes and tools Working software over comprehensive documentation Customer collaboration over contract negotiation Responding to change over following a plan That is, while there is value in the items on the right, we have the items on the left more." [agilealliance.org] The twelve principles of agile software development (Agile Alliance 2001) are:

1) The highest priority is to satisfy the customer through early and continuous delivery of valuable software2) the welcoming of changing requirements, even development, for the benefit of the customer's competitive advantage,3) frequent delivery of working software, the release cycle ranging from a couple of weeks to a couple of months, with a preference for a shorter timescale,4) daily collaboration of business people and developers throughout the project,5) building of projects around motivated individuals by offering them an appropriate environment and the support they need, and trusting them to get the job done, 6) emphasis on face-to-face conversation for conveying information and within a development team, 7) working software is the primary measure of progress, 8) agile processes promote a sustainable development pace for sponsors, developers, and users, 9) continuous attention to technical excellence and good design enhances agility, 10) simplicity is essential for maximising the amount of work not having to be done, 11) self-organising teams give best results in terms of architectures, requirements, and designs, 12) regular reflection of teams on how to become more effective, and tuning and adjusting its behaviour accordingly.

These days extensive research is being carried out to get best of agile development as[9] shows in a tree that no agile process follows all the principles. Lean software development has five bottlenecks, XP itself has two, SCRUM has two and FDD also has seven bottlenecks.

As many authors say that agile development becomes industry standard but Agile processes also have limitations as[8]discusses the limitations of agile on its 11 assumptions which says none of the agile processes is a silver bullet to fit all these assumptions.

To get best traditional approach and agile approach has to combine. [7] says that the companies quite expertly combine agile and traditional practices and adjust their practices according to the situation at hand. Figure 2 shows the effects (benefits or drawbacks) in both the methods and also insists on the cumulative methods development since there has been a movement from no methods, via tradidional method to agile method.

Research is going on to combine agile with other processes , models .One is in [11] which concludes through a case study that SPLE(software product line engineering) and agile software development are complementary to each other.

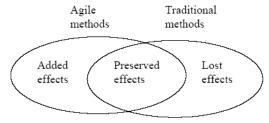


Figure 2: Relation between identified effects [14]

Step 1.	Rate the project's environmental, agile, and plan- driven risks. If uncertain about ratings, buy information via prototyping, data collection, and analysis.
Step 2a.	If agility risks dominate plan-driven risks, go Risk- based Plan-driven.
Step 2b.	If plan-driven risks dominate agility risks, go Risk- based Agile.
Step 3.	If parts of the application satisfy 2a and others 2b, architect the application to encapsulate the agile parts. Go Risk-based Agile in the agile parts, and Risk-based Plan-driven elsewhere.
Step 4.	Establish an overall project strategy by integrating individual risk mitigation plans
Step 5.	Monitor progress and risks/opportunities, readjust balance and process as appropriate.

Table 1: Summary of Risk based approach.[15]

Table 1 shows a risk based approach to develop a balanced development strategy. [16] discusses about the process which appears to be generic i.e. amenable to use for building any type of system, including web applications; in a context where risk analysis is important.

One key research area related to agile processes is in software process improvement. [18] reflects such a need as in Table 2 the differences between traditional and agile software development approaches gives an iterative process improvement technique as a solution with five case studies.

	Traditional software development and SPI	Agile software development and SPI
Software development process	Universal approach and repeatable solution to provide predictability and high assurance	Flexible approach adapted with collective understanding of contextual needs to provide faster development times, responsiveness to rapid changes, increased customer satisfaction, and lower defect rates.
Process control	Control on organizational level	Self-organizing teams
Primary means of knowledge transfer	Document based knowledge transfer	Face-to-face communication
Immediate focus of process improvement	Improvement of organizational software development processes/(future projects)	Improvement of daily working practices of ongoing project

Table 2: Underlying differences of traditional and agile software development and SPI [18]

[10] shows how the CMMI could be used in assessing agile software development or in the situation where organization is planning to change its process towards agility. Following Table 3 concludes that "While CMMI

creates an organizational discipline; XP eases the daily life by providing pragmatic, end-result-oriented practices. CMMI and XP can be used together very well and their synergy is very strong."

	Cards	Pair	First Test
		Programming	then
			Coding
Requirements	++	N/A	+++
Management			
Project	+++	++	++
Planning			
Configuratio		N/A	N/A
n			
Management			

Table 3: Relationship between some of the CMMI process and some XP practices [17].

Among the limitations of agile methods mentioned in [12] one is the lack of attention to design and architectural issues. Boehm has done great work on architecture and pointed out that there is a risk of architectural mistakes that cannot be detected easily by external reviewers due to lack of documentation in agile development.

Software Architecture and Agile Software Development:

Software architecture of a program or computing system is the structure or structures of the system which comprise software elements , the externally visible properties of those elements and the relationships among them. [23]

To accommodate changes at any level of development results in compromise on quality in lightweight processes. Moreover agile development produces specified products. Our interest is in how much agile can contribute to produce generalized products, reusable artifacts. [20] maintains that these two (agile approaches and software architecture) seemingly opposing views to software engineering can be integrated but it requires that experts from both fields work together to overcome evident challenges in bridging these two paradigms together and insists on the need of research on integrating architecture-centric methods in agile approaches.

In architecture oriented agile software development, the main considerations are in which iteration the architecture will be designed, how much extra cost has to bear, is customer interested in architecture development. Future research is open in architecture oriented agile software development to get answers to these questions so as we can get more quality product and more productivity.

Moreover, Software architecture research aims at

reducing development costs by identifying communalities among closely related products. Software architecture entails the principal design decisions concerning the system and is rather orthogonal to the development process[2]. Architect has to detect nonfunctional requirements from the requirements stated by the customer. He has to work with His vision. [23] shows there could be quality attribute trade —offs which should be taken care off. It is also point of consideration that in agile development the architect is one of the developer's team or an individual one.

Though introduction of architecture reduces many risks but it also introduces many risks called architectural risks. [21] Summarizes that architectural risks in agile processes can be handled by two ways. Architectural risks that we know in advance can be handled in a time boxed iteration zero, where no features are planned to delivered. Small architectural risks can be handled as they arise during iterations, but large architectural risks must be promoted to be on par with features, and inserted into a combined feature and risk backlog.

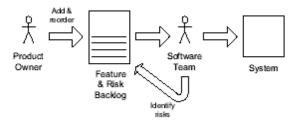


Figure 3: Feature and Risk Backlog based ASD[21]

One way to incorporate risk into an agile process is to convert the feature backlog into a feature and risk backlog. The product owner adds features and the software teams add risks. The software team must help the product owner to understand the technical risks and simply prioritize the backlog [21].

Architectural Patterns: An architectural pattern is a description of element and relation types together with a set of constraints on how they may be used.[23] Some authors specify the patterns what D.Garlan and His collaborators call styles. In most cases arch patterns are considered in close connection with object orientation. Object oriented language constructs like abstract classes or inheritance, which support the architectural pattern idea in a very elegant way.[4]

Pattern is being decided on the non-functional requirements of the product. Single pattern or combination of patterns is being used to design the architecture by the architect by keeping in mind the hindrance to each of the non-functional properties because of one another.

As patterns are already fully tested and can be easily adapted ,enhances the reusability . One negative point

attached to it is that Pattern reuse depends upon several factors.[2]

From the software evaluation point of view, if the abstract scenarios that characterize the quality attributes satisfied by the patterns used in the software are available, it will improve software architecture evaluation and reduce the time and resources required to gather scenarios from scratch for each evaluation effort.[3]

Reusability: Number of techniques are available to support reusability. Considerable research and development is going on in reuse; industry standards like CORBA have been created for component interaction; and many domain specific architecture, toolkits, application generators and other related products that support reuse and open systems have been developed[1]. Architecture reusability can be increased by defining levels as in FIM architecture [5] which operates at three different levels of reuse: Federation, domain and application.

Agile software development and reusability in the software engineering lays the same foundation of the quality product and reduction in development time. To get a generalized product through agile processes software architecture has been introduced which supports modifiability also. A repository can be built to place various artifacts like patterns, components, and reference architectures in ASD.

Focus on how a non-functional property reusability relates to the software architecture of a system [13]. [22] suggested software process model for reuse based software development approach.

Conclusion: Reusability reduces the complexity of design process. Introduction of software architecture reduces the risks and increases the modifiability at later stages even. It is concluded that agile development which has promising future in the software industry and can fulfill the demands of the industry can be improved more. Adding a slight touch of traditional approach, Object oriented patterns for reusability at design, code and test level, and architecture specific designs will definitely make space for reusability and reusable artifacts. Architecture oriented agile development is an open research area. Designing and developing specialized cum generalized components can be achieved by using object oriented patterns that is a new dimension to be explored more for effective ASD.

Bibliography:

- [1] Garlen David, A. Robert, and ockerbloom john, nov. 1995, Architectural mismatch: Why reuse is so hard Vol. 12, no 6.
- [2] Cimpan Sorana , Couturier Vincent, 2008, Can styles improve architectural pattern reuse? Proceedings of 7th

- working IEEE/IFIP conference on software architeture (WICSA 2008) 263-266, 2008, ISBN: 978-0-7695-3092-5
- [3] Babar Ali M., Improving the reuse of pattern-based knowledge in software architecting ,www.patternforge.net/wiki/images/3/35/alibabar.pdf.
- [4] A. Marco Components , connectors and architectural patterns, www.citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.12.8257&rep=rep1 & type=pdf
- [5] Gomma H., Farukh A.G.,1999, A reusable architecture for federated client/server systems, Proceedings of the 1999 symposium on software reusability, los angeles, California, US, pages 113-121,ISBN: 1-58113-101-1
- [6] Salo Outi, Enabling software process improvement in agile software development teams and organizations, VIT Publications 618
- [7] H. Christina ,D. Yvonne, G. Bjorn ,and Z. Stefan ,2006, How agile are industrial software development practices?,The journal of systems and software 79(2006) 1295-1311.
- [8] Turk Dan, F.Robert ,and R. Bernhard,may 2002,Limitations of agile software processes ,3rd international conference on XP and agile processes in software engineering(XP 2002).
- [9] M. Asta, A. Vaidos, 2008, Bottlenecks in agile software development using theory of constraints(TOC) principles, Gothenburg, Sweden 2008.
- [10] P.Minna, and M. Annukka, 2006, An approach for using CMMI in agile software development assessments: experiences from three case studies, SPICE 2006.
- [11] Geir K. Hanssen , Tor E. Faegri, 2008,Process Fusion : An industrial case study on agile software product line engineering, The journal of systems and software 81(2008) 843-854.
- [12] Tore Dyba, Torgeir Dingsoyr,2008, Empirical studies and agile software development: A systematic review. Information and software technology 50(2008) 833-859
- [13] Frances Paulisch , Siemens AG,1-2 nov., 1994. Software architecture nad Reuse —an inherent conflict?,3rd international conference onsoftwrae reuse, page 214.
- [14] C. Stefan, 2008, Using Agile Methods? Expected effects ,17th International conference on information systems development (ISD 2008).Paphas, Cyprus ,aug 25-27.2008.
- [15] B. Barry, T. Richard, 2004, Balancing Agility and Discipline: Evaluating and Integrating Agile and Plan-Driven methods, ICSE 2004, Pp. 718-719
- [16] Xiaocheng Ge, Richard F. Paige, Fiona A.C. Polock, Howard Chivers, Phillip J. Brooke, 2006, Agile development of secure web applications, Proceedings of the 6th international conference on web engg., pages 305-312.ISBN: 1-59593-352-2
- [17] Orhan Kalayci, Nitelik Danismanlik ltd.,Sait Dinmez,Emel Saygin,Serden Ferhatoglu, gulfer Akgun,Senol Bolat,Hasan Ozkeser, BIMAR Bilgi islem Hizmetleri A.S., Real Life Experience Using CMMI L2

- processes and XP Practices, www.nitelik.net/yayinlar/PSQT/internalpilot.pdf
- [18] Salo Outi, and A. Pekka, An Iterative improvement process for Agile software development, www.agileitea.org/public/papers/SPIP.pdf
- [19] Michael Cusumano, Alan MacCormack, Chris F. Kemerer, Willliam Crandall, june 17, 2003, A Global Survey of Software development Practices, Version 3.1, Forthcoming IEEE software
- [20]Babar Ali M., and A. Pekka,29th dec.,2009, Architecture –Centric method and agile approaches,10th international conference on agile processes.
- [21] F. George, The risk -centric model for software architecture
- ,www.mysite.verizon.net/dennis.mancl/oopsla09/risk-centric-software-architecture-positio-paper.pdf
- [22] Jasmine K.S., Dr. R. Vasantha, july 2-4,2008, A new process model for reuse based Software development approach, Proceedings of the world congress on engineering 2008 vol. 1, WCE 2008, London U.K.
- [23] Softwarre architecture in practice ,second edition , Len Bass ,Paul Clements, Rick Kazman

Aggregating Intrusion Detection System Alerts Based on *Row Echelon Form* Concept

Homam El-Taj, Omar Abouabdalla, Ahmed Manasrah, Moein Mayeh, Mohammed Elhalabi

> National Advanced IPv6 Center (NAv6) UNIVERSITI SAINS MALAYSIA Penang, Malaysia 11800

Abstract— Intrusion Detection Systems (IDS) are one of the well-known systems used to secure the computer environments, these systems triggers thousands of alerts per day to become a serious issue to the analyst, because they need to analyze the severity of the alerts and other issues such as the IP addresses, ports and so on to get better understanding about the relations between the alerts. This will lead to have a better understanding about the attacks. This paper Investigates the most popular aggregation methods, which deals with IDS alerts. In addition, we propose Time Threshold Aggregation algorithm (TTA) to handle IDS alerts. TTA is based on time as a main component to aggregate the alerts. On the other hand, TTA supports aggregating alerts without threshold, which can be done by setting the threshold value to 0.

Keywords—Intrusion Detection System, False Positive, Redundant Alerts, Alert Aggregation.

I. INTRODUCTION

The reason behind creating intrusion detection systems (IDS) is because of the huge amount of threats and attacks over the internet and wide networks. In the other, hand IDS triggers huge amount of alerts because of these threats; therefore managing and controlling these alerts need to be studied which led the researchers to investigate these alerts to create methods and techniques such as aggregation to minimize the amount of alerts and group them to make them fewer and to reduce the analyzing process time. Such a progress like this directed to minimize the false positive of IDS too. A good knowledge of IDS and their alerts should be known for better understanding of the aggregation technique.

A. Intrusion Detection System (IDS)

IDS as a system triggers alert or a group of alerts if there is an intrusion of the monitored network based on analyzing the activities, these activities are collected from the network packets stream. IDS has two ways of detecting intrusions either by using *anomaly* [1] technique or *misuse* technique [2] or by merging both techniques starts by checking whether the attack signature saved in the database as a misuse technique then apply the anomaly techniques to check if it is

anomaly attack. *Misuse* detecting techniques look for a malicious signature or pattern of the threat based on a set of rules or signatures to detect intrusive behavior while *anomaly* detection technique determines the abnormality of network flow by measuring the distance between the suspicious activities and the norm based on a chosen threshold. The main differences between these two techniques are based on detecting the novel attacks and the false positive rate, where *anomaly* techniques can detect novel attacks and they have a high rate of false positive, *misuse* techniques in the other hand have low rate of false positive without the ability of detecting novel attacks. To differentiate between these two techniques and have a better background you may refer to [3-6].

B. IDS Standard Alerts Format

There is a variety on the sensor types, these sensors trigger a non standard formats of alerts, which led to create the standardization format. One of these standards is the Intrusion Detection Message Exchange Format (IDMEF). This standard was built with Extensible Markup Language (XML) and it has the flexibility to accommodate different needs [7].

II. AGGREGATION TECHNIQUES

Aggregation technique is one the major parts of IDS studies for grouping and minimizing the alerts to ease the process of analyzing them by removing the redundant alerts. Aggregation techniques group the IDS alerts based on the similarity of the alert features, since some of the alerts related to one event usually they have similar features, so they will be aggregated into one alert. This paper will try to give the answers of the following questions: how to define the alert features? How to calculate the similarity of them?

Valdes [8] proposed an aggregation algorithm by including the five features: source IP addresses, source ports, destination IP addresses, source ports and alert generation time. The compression result of each feature is a value between 0 and 1, while the similarity calculation and the weights of each feature depend on predefined values. But the

researcher didn't mention the method of defined the similarity and the weights values. Another proposed solution by [9] was based on the exact matching which gives us the result of 0 or 1, so this algorithm is weak because it reduces a little amount of alerts. Another approach based o [8] algorithm's done by [10] give us slightly different experiment results because he used only the source IP addresses and alert generation time.

Different aggregation technique is introduced by [11], this technique aggregates the alerts by categorizing their features into four classes, then a similarity operator used to compute the similarity of the same features class, but there is no discussion on the computation methods for each features class. Oliver [12] suggested that the alerts should be categorized by attack intensions basis, using the subsequent aggregation processes. Investigating the ideas from the previous proposed methods leads to this proposed algorithm. Time threshold aggregation algorithm (TTA) works on extracting the features' alerts to categories them into groups based on the similarity on these features maintaining the integrity of the alerts' features.

III. PROPOSED ALGORITHM

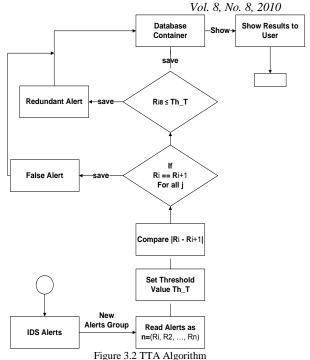
As mentioned in previous sections, based on some of alerts' features or all of them, the researchers are trying to make their aggregation algorithms without any consideration on the alerts' trigger time. In this proposed algorithm, the merging of any two alerts or more will be based on a threshold value, which should give more accuracy combination results. Figure 3.1 explain the TTA algorithm.

I. Time Threshold Aggregation algorithm

TTA works as illustrated in following:

- (1) Read IDS alert as $n = (R_i, R_2, ..., R_n)$
- (2) Get the first row items as R_i where $i = \{j_1, j_2,..., j_8\}$
- (3) Set the Threshold *Th T* value
- (4) Iteration I = n-1
- (5) Compare the Rows by $|R_i R_{i+1}|$
- (6) Update the R_{i_8} Value $R_{i_8} = R_{i+1_8} I \check{f} R_{i+1_1}, R_{i+1_2}, \dots, R_{i+1_8} = 0$ (7) While I \geq 1 Do
- (8) Delete R_{i+1} if $(R_{i+1_i} \text{ for } i_1,..., i_7 = 0 \& |R_{i_8}| \le Th_T)$
- (9) I = I-1

TTA based on the Row Echelon Form [13] Concept, in TTA we conseder the redundunt alerts as false positive alerts if there is an exact matching $(R_i = R_{i+1})$ for $i_1, ..., i_8$, and if the different is only in the time feature is repeated i₈ we conseder it as a real alert.



To understand the algorithm of ATT, check the following Example:

(1) Let the sample of the A took from the table 3.1 and the Th T = 2. From table 3.1 we get $A_1 = \{4,$ 1, 2, 2, 3, 1, 2, 1, $A_2 = \{4, 1, 2, 2, 3, 1, 2, 2\}$, $A_3 =$ $\{4, 2, 1, 2, 3, 1, 3, 2\}, ..., A_{13}\{4, 1, 2, 2, 3, 1, 2, 9\}.$

Table 3.1 Example of A								
4	1	2	2	3	1	2	1	
4	1	2	2	3	1	2	2	
4	2	1	2	3	1	3	2	
4	1	2	2	3	1	2	3	
7	1	4	2	1	1	1	3	
4	1	2	2	3	1	2	3	
7	1	4	2	1	1	1	4	
4	1	2	2	3	1	2	5	
7	3	4	6	1	1	1	6	
7	1	4	2	1	1	1	6	
4	1	2	2	3	1	2	7	
7	1	4	2	1	1	1	8	
4	1	2	2	3	1	2	9	

(2) We Multiply $(A_2, ..., A_{13})^*$ -1 then do apply

$$A_{i+1}^{13} = |A_i - A_{i+1}^{13}| \tag{1}$$

2, -2, -3, -1, -2,-9}. After we apply equation 1 the result will be like this $A_1 = \{4, 1, 2, 2, 3, 1, 2, 1\},\$

$$A_2 = \{0, 0, 0, 0, 0, 0, 0, 1\}, A_3 = \{0, 1, 1, 0, 0, 0, 0, 2\}, ..., A_{13}\{0, 0, 0, 0, 0, 0, 0, 3\}.$$

(3) After each time we apply equation 1 we check if A_{i_8} need to be updated by

$$A_{i_0} = A_{i+1_0} \tag{2}$$

- (4) Equation 2 will be used only $A_{i+1_1}, A_{i+1_2}, \dots, A_{i+1_8} = 0$, this equation can be use in the case of A_2 , A_4 , A_8 , A_{11} as $\{2, 3, 5, 6\}$ in the case of A_{13} A_{13_8} has not been updated because $A_{13_8} > Th_T$.
- (4) We eliminate the zero's rows regarding $i_1, ..., i_7$ to get a new set of A as table 3.2

Table 3.2 New set of A							
4	1	2	2	3	1	2	1
4	2	1	2	3	1	3	2
4	2	1	2	3	1	3	2
7	1	4	2	1	1	1	3
7	1	4	2	1	1	1	4
7	3	4	6	1	1	1	6
7	1	4	2	1	1	1	6
7	1	4	2	1	1	1	8
4	1	2	2	3	1	2	9

(5) We repeat step (1, 2, 3, 4, 5) until there is no rows left.

II. Mathematically Proof:

If we consider A as the alerts' dataset then

$$A = \begin{pmatrix} 4 & 1 & 2 & 2 & 3 & 1 & 2 & 1 \\ 4 & 1 & 2 & 2 & 3 & 1 & 2 & 2 \\ 4 & 2 & 1 & 2 & 3 & 1 & 3 & 2 \\ 4 & 1 & 2 & 2 & 3 & 1 & 2 & 3 \\ 7 & 1 & 4 & 2 & 1 & 1 & 1 & 3 \\ 4 & 1 & 2 & 2 & 3 & 1 & 2 & 3 \\ 7 & 1 & 4 & 2 & 1 & 1 & 1 & 4 \\ 4 & 1 & 2 & 2 & 3 & 1 & 2 & 5 \end{pmatrix} \Rightarrow A = \begin{bmatrix} B \\ \ddot{N} \end{bmatrix}$$

Where B is the set of alerts related to the hyper alerts (Produced alerts from A) and N is the set of alerts representing the false positive alerts (None related).

From the first iteration we got $i_1 = [0\ 0\ 0\ 0\ 0\ 0\ 0]$ and from second iteration we got $n_1 = [0\ 1\ -1\ 0\ 0\ 0\ 1\ 1]$ therefore:

$$\therefore B = \begin{bmatrix} i_1 \\ i_2 \\ i_3 \\ i_4 \end{bmatrix} = \begin{bmatrix} 4 & 1 & 2 & 2 & 1 & 3 & 2 & 1 \\ 4 & 1 & 2 & 2 & 1 & 3 & 2 & 2 \\ 4 & 1 & 2 & 2 & 1 & 3 & 2 & 3 \\ 4 & 1 & 2 & 2 & 1 & 3 & 2 & 3 \\ 4 & 1 & 2 & 2 & 1 & 3 & 2 & 5 \end{bmatrix}$$
 Al

erts
related
1

And
$$N = \begin{bmatrix} n_1 \\ n_2 \\ n_3 \end{bmatrix} = \begin{bmatrix} 4 & 2 & 1 & 2 & 3 & 1 & 3 & 2 \\ 7 & 1 & 4 & 2 & 1 & 1 & 1 & 3 \\ 7 & 1 & 4 & 2 & 1 & 1 & 1 & 4 \end{bmatrix}$$

We use the set final set of N to the new alerts' dataset, and we keep repeating the algorithm steps until we get empty N set and $B = [i_1] = [7 \ 1 \ 4 \ 2 \ 1 \ 1 \ 3]$

The final aggregated file Agg will be like follow

$$Agg = \begin{bmatrix} 4 & 1 & 2 & 2 & 3 & 1 & 2 & 1 \\ 4 & 2 & 1 & 2 & 3 & 1 & 3 & 2 \\ 7 & 1 & 4 & 2 & 1 & 1 & 1 & 3 \end{bmatrix}$$

Table 4.2 Aggregated Alerts Agg

4	1	2	2	3	1	2	1
4	2	1	2	3	1	3	2
7	1	4	2	1	1	1	3

III. Using Time Threshold Aggregation algorithm on IDS Alerts

As mentioned in section 1 and 2, alerts contains many features in TTA we focus in 8 features to do the aggregation (Source IP, destination IP, Source port, destination port, Severity, Protocol, Alert Classification and Time) so if we took each alert as a row in our algorithm it can give a promising results.

IV. USING TIME AS A MAIN FEATURE

Most of the previous studies didn't take the alert trigger time as one of the extracted features. We believe the time threshold will effects the accuracy of the aggregation result by taking it as one of the aggregation features, based on the alert trigger time, the process of analyzing the alerts will be easier. The analysts would like to know the severity of the alerts based on the amount of the alerts by the same features which this algorithm can show. Based on the threshold Th_T that the user will select; the amount of the aggregated alerts will be changed.

V. DISCUSSION

This paper presented the ATT algorithm for aggregating alerts from any intrusion detection systems. The main advantage of the proposed framework is to improve the alert aggregation process especially when it is related to triggered alert time. The advantages of ATT are: to minimize the amount of alerts, remove the redundant alerts and to remove the false alerts.

Other benefits of the proposed algorithm are: Firstly, to obtain the most benefit from the alerts by making use of the supporting features from the alerts itself which is controlled by the user. Secondly, by analyzing any type of alerts in a standard format, the algorithm provides flexibility to make use of the enriched alerts for aggregating purpose rather than any complicated techniques. Thirdly, this algorithm can ease the study of the alerts severity when they can be related to the aggregated alert groups. Finally, using this algorithm will give us accurate and less number of alerts since it is based on the threshold value which can be modified by increasing or decreasing it. By setting the threshold value to 0 the alerts will be aggregated by exact matching. In other words, the aggregated alerts should be approximately 0 aggregation with the same amount of output alerts, and since it is very hard that two alerts will be triggered in the same time from the same sensor or the same IDS, the amount of the aggregated alerts is high. By increasing the value of the threshold the amount of output alerts will be decreased. After a number of trials the user can tell what are the right value of threshold should be.

VI. CONCLUSION AND FUTURE WORK

TTA can be used as an aggregation method to any file containing a group of Items with extracted features. In the stage of programming TTA, it should give the user the ability to choose the number of extracted features from the IDS alerts. TTA can be implemented in using parallel technique for the comparison part between the alerts to give a better time results.

ACKNOWLEDGMENT

This research was supported by the National Advanced IPv6 Center (NAv6) in UNIVERSITI SAINS MALAYSIA (USM).

REFERENCES

[1] W. Fan, M. Miller, S. Stolfo, W. Lee, and P. Chan, "Using artificial anomalies to detect unknown and known network intrusions," *Knowledge and Information Systems*, vol. 6, pp. 507-527, 2004.

- [2] M. Sheikhan and Z. Jadidi, "Misuse Detection Using Hybrid of Association Rule Mining and Connectionist Modeling," *World Applied Sciences I*, vol. 7, pp. 31-37, 2009.
- [3] Y. Liao and V. R. Vemuri, "Use of K-Nearest Neighbor classifier for intrusion detection," *Computers & Security*, vol. 21, pp. 439-448, 2002.
- [4] A. Alharby and H. Imai, "IDS false alarm reduction using continuous and discontinuous patterns," Springer, 2005, pp. 192-205.
- [5] A. Sundaram, "An introduction to intrusion detection," *Crossroads*, vol. 2, pp. 3-7, 1996.
- [6] M. J. Ranum, "False Positives: A User's Guide to Making Sense of IDS Alerts," in http://searchsecurity.techtarget.com/whitepaperPage/0,293857,sid14 gci903698,00.html, I. L. IDSC, Ed., 2003.
- [7] H. Debar, D. Curry, and B. Feinstein, "The Intrusion Detection Message Exchange Format (IDMEF)." vol. 2010: March 2007, 2007.
- [8] A. Valdes and K. Skinner, "Probabilistic alert correlation," in the Fourth International Symposium on Recent Advances in Intrusion Detection, 2001, pp. 54–68.
- [9] H. Debar and A. Wespi, "Aggregation and correlation of intrusion-detection alerts," in 4th International Symposium on Recent Advance in Intrusion Detection(RAID) 2001, 2001, pp. 85-103.
- [10] C. Mu, H. Huang, S. Tian, Y. Lin, and Y. Qin, "Intrusion-detection alerts processing based on fuzzy comprehensive evaluation," *Jisuanji Yanjiu yu Fazhan(Computer Research and Development)*, vol. 42, pp. 1679-1685, 2005.
- [11] F. Autrel and F. Cuppens, "Using an intrusion detection alert similarity operator to aggregate and fuse alerts" in *The 4th Conference on Security and Network Architecture* Batz sur Mer, France, 2005.
- [12] O. Dain and R. K. Cunningham, "Fusing a heterogeneous alert stream into scenarios," *Applications of Data Mining and Computer Security*, 2002.
- [13] J. Faugère, "A new efficient algorithm for computing Grobner bases (F4)," *Journal of Pure and Applied Algebra.*, vol. 139 pp. 61-88, 1999.

Evaluation of Vision based Surface Roughness using Wavelet Transforms with Neural Network Approach

*T.K.Thivakaran *Research Scholar, MS University, Tirunelyeli – 627012.INDIA

**Professor, Department of CSE, Annamalai University, Chidambaram – 620 024.INDIA

Abstract---Machine vision for industry has generated a great deal of interest in the technical community over the past several years. Extensive research has been performed on machine vision applications in manufacturing, because it has the advantage of being non-contact and as well faster than the contact methods. Using Machine Vision, it is possible to evaluate and analyze the area of the surface, in which machine vision extracted the information with the help of array of sensors to enable the user to make intelligent decision based on the applications. In this work, Estimation of surface roughness has been done and analyzed using digital images of machined surface obtained by Machine vision system. Features are extracted from the enhanced images in spatial frequency domain using a two dimensional Fourier Transform and Wavelet Transform. An artificial neural network (ANN) is trained using feature extracted values as input obtained from wavelet Transform and tested to get Rt as output. The estimated roughness parameter (Rt) results based on ANN is compared with the R_t values obtained from Stylus method and the best correlation between both the values are determined.

Keywords--- Surface roughness, Machine vision, Milling, Grinding, Wavelet Transform, Neural Network.

I. INTRODUCTION

The quality of components produced is of main concern to the manufacturing industry, which normally refers to dimensional accuracy, form and surface finish. Therefore, the inspection of surface roughness of the work piece is very important to assess the quality of a component, which is normally performed using stylus type devices, which correlate the vertical displacement of a diamond-tipped stylus to the roughness of the surface under investigation. But, the limitations of stylus techniques have already been reported in detail in [6, 5, 4]. Machine Vision typically employs a camera, a frame grabber, a digitizer and a processor for inspection tasks where precision, repetition and/or high speed are needed. The histograms of the surface image have been utilized to characterize surface roughness and quality. Fourier transform (FT) of the digitized surface image in which the magnitude and frequency information obtained from the FT are used as measurement parameters of the surface finish. These methods use the basic assumption that the surface of the specimen is completely flat and there is no inclination when the images are captured. Even a small inclination of the specimen may result in inconsistent estimation of roughness of components using machine vision primarily due to the fact that illumination, shadow on the images is likely to be different.

In this work, the machined surfaces are captured using a Machine Vision system. Following the image enhancement, the features are extracted and then the roughness parameters are estimated and analyzed. Here wavelet is used to extract the features of the enhanced image, and an artificial neural network (ANN) is developed to predict the surface roughness. The results are compared with that obtained using the standard stylus method.

II. ROUGHNESS PARAMETERS

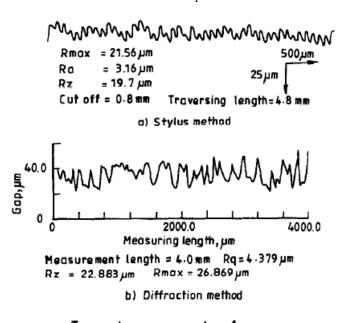
The machined surfaces are generally characterized by three kinds of errors (i) form errors, (ii) waviness, and (iii) surface roughness. The concept of roughness is often described with terms such as 'uneven',' irregular', 'coarse in texture', broken by prominences', and other similar ones (Thomas, 1999). Similar to some surface properties such as hardness, the value of surface roughness depends on the scale of measurement. In addition, the concept roughness has statistical implications as it considers factors such as sample size and sampling interval. The one parameter that is standardized all over the world and is specified and measured far more frequently than any other is the arithmetic average roughness height, or Roughness Average. Universally called Ra, it was formerly known as AA (Arithmetic Average) in the United States and CLA (Center Line Average) in the United Kingdom. It is defined as the arithmetic mean of the departures of the profile from the mean line.

Rq (or also known as RMS) is the root mean-square average of the departures of the roughness profile from the mean line. Rq has statistical significance because it represents the standard deviation of the profile heights and it is used in the more complex computation of skewness, the measure of the symmetry of a profile about the mean line.

$$R_{q} = \left\{ \frac{1}{L} \int_{0}^{L} z^{2} dx \right\}^{1/2} \dots (1)$$

A. Profiles for Turned Machined Components

Figure 1(a) and figure 1(c) shows the profiles obtained for a turned component with a stylus instrument. Similarly, figure 1(b) and 1(d) shows the gap profiles obtained for the same turned components by diffraction method. In both graphs, 'z' is the deviation of the points on the profile from the mean-line. It can be observed that appreciable differences in the diffraction pattern are seen for large variations in the gap and therefore good comparison of results is guaranteed in both only for turned components of medium roughness. For very rough surfaces scattering is observed. A limitation in the usage of the different methods is that the smoothness of the edge plays a crucial role in the evaluation of the finish of the components.



Turned component - 1

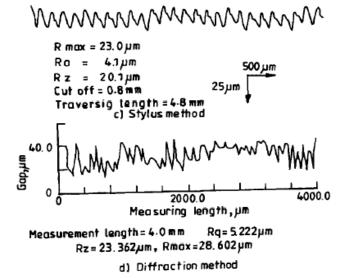


Fig. 1(a), (c) The profiles obtained for a turned component with a stylus instrument. (b), (d) the gap profiles obtained for the same turned components by diffraction method.

B. Profile for Ground Machined Components

Figure 2(a) shows the profiles obtained for a ground component with a stylus instrument. Similarly, figure 3(b) shows the gap profiles obtained for the same ground components by diffraction method. In both graphs, 'z' is the deviation of the points on the profile from the mean-line.

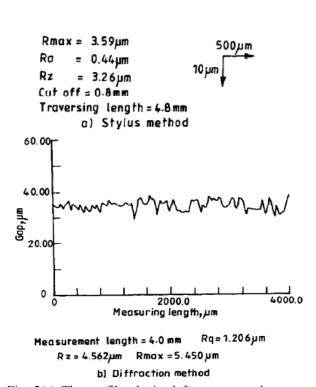


Fig. 2(a) The profile obtained from a ground component with a stylus instrument, (b) The gap profile obtained for the same ground component by diffraction method.

III. SPECTRUM TECHNIQUES FOR FEATURE EXTRACTION

A. Fourier spectrum

The Fourier spectrum is the frequency domain counterpart of the autocorrelation function. The FT of the correlation is used, which corresponds to the power spectral density function and describe how the power in a signal is distributed over frequency. The power spectrum can reveal the presence of offset, or periodic structures in a data set.

B. Wavelet Transform (WT)

The wavelet is a tool in surface texture analysis and can decompose a surface into multi-scale representation in a very efficient way. The wavelet transform (WT) is a mapping of the signal to the time-scale joint representation. By WT, the decomposition of a signal with a real

orthonormal bases $\Psi_{mn}(x)$ obtained through translation and dilation of a kernel function $\Psi(x)$ known as mother wavelet as given in eqn. [2],

$$\psi_{m,n}(x) = 2^{-m/2} \psi(2^{-m} x - n) \qquad \dots (2)$$

Where, m,n are integers. To construct the mother wavelet $\Psi(x)$, it is required to determine a scaling function $\phi(x)$ given in eqn. [3],

$$\phi(x) = \sqrt{2} \sum_{k} h(k)\phi(2x - k)$$
 ... (3)

Then, the mother wavelet $\mathcal{Y}(x)$ is related to the scaling function as in eqn. [4],

$$\psi(x) = \sqrt{2} \sum_{k} g(k)\phi(2x - k) \qquad \dots (4)$$

where

$$g(k) = (-1)^k h(1-k)$$

The coefficients h(k) have to meet several conditions for the set of basis wavelet functions to be unique, be orthonormal and also have a certain degree of regularity.

C. Wavelet Transform for Signals

In two dimensional cases, the one dimensional wavelet transforms are applied along both the horizontal and vertical directions $\phi(x)$ is a one dimensional real, sequence integral scaling function defined as in [5]

$$\phi_{j,k}(x) = 2^{\frac{j}{2}} \phi(2^{j} x - k)$$
 ... (5)

Translation k determines the position of this one dimensional function along the x- axis, scale j determine its width along x axis and $2^{\frac{j}{2}}$ controls its height and amplitude. This one dimensional scaling function satisfies these conditions:

- $\phi_{i,k}$ is orthogonal to its integer translates.
- The set of functions that can be represented as a series expansion of $\phi_{j,k}$ at low scale is contained within those at higher scale.

So, the difference between any two sets of $\phi_{j,k}$ is represented by a companion wavelet function $\psi_{j,k}$ defined

in eqn. [6],
$$\psi_{j,k}(x) = 2^{j/2} \psi(2^j x - k)$$

... (6)

Then, the 2 dimensional DWT functions are the linear products of scaling and wavelet functions $\phi(x)$ and $\psi(x)$ yielding the eqn. [7] through eqn. [9].

$$\psi^{H}(x,y) = \psi(x).\phi(y) \qquad \dots (7)$$

$$\psi^{V}(x,y) = \phi(x).\psi(y) \qquad \dots (8)$$

$$\psi^{D}(x, y) = \psi(x).\psi(y)$$
 ... (9)

where, $\psi^H(x,y)$, $\psi^V(x,y)$ and $\psi^D(x,y)$ are called the horizontal, vertical and diagonal wavelets. Thus, DWT is well localized and allows decomposition in three directions, namely, horizontal, vertical and diagonal respectively.

D. Features of Wavelets

In this application, the features are extracted using a wavelet which belongs to a family of orthogonal wavelets. The mother wavelet (DB4), its corresponding scaling and wavelet functions and the decomposition filters are shown in Figure 3 and Figure 4 respectively.

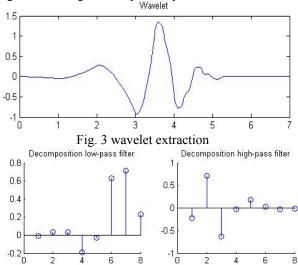


Fig. 4 Decomposition of low-pass filter $h_{\varphi}(-n)$ and highpass filter $h_{\psi}(-m)$

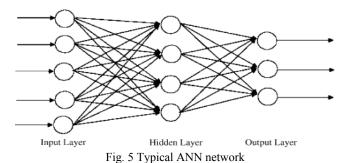
The DB4 scaling function is given by $a_i = h_0 s_{2i} + h_1 s_{2i+1} + h_2 s_{2i+2} + h_3 s_{2i+3} \qquad \dots (10)$ $a[i] = h_0 s[2i] + h_1 s[2i+1] + h_2 s[2i+2] + h_3 s[2i+3] \qquad \dots (11)$

The Daubechies DB4 wavelet function is given by $c_i = g_0 s_{2i} + g_1 s_{2i+1} + g_2 s_{2i+2} + g_3 s_{2i+3} \qquad \dots (12)$ $c[i] = g_0 s[2i] + g_1 s[2i+1] + g_2 s[2i+2] + g_3 s[2i+3] \qquad \dots (13)$

IV. NEURAL NETWORKS FOR SURFACE ROUGHNESS ASSESSMENT

The roughness features extracted from the machined images, are fed as input to an ANN to predict the roughness value R_t. ANN consists of a number of elementary units called neurons. A neuron is a simple processor, which can take multiple inputs and produce an output. Each input into the neuron has an associated weight that determines the "intensity" of the input. The processes that a neuron performs are: multiplication of each of the inputs by its respective weight, adding up the resulting numbers for all the inputs and determination of the output according to the result of this summation and an activation function. Data is fed into the network through an input layer, it is processed through one or more intermediate hidden layers and finally

fed out of the network through an output layer as shown in Figure 5.



V. PROPOSED SYSTEM FOR SURFACE ROUGHNESS EVALUATION

The methodology and block diagram of proposed Machine vision system is shown in Figure 6(a) and Figure 6(b).

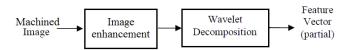


Fig.6 (a) Block diagram of proposed system

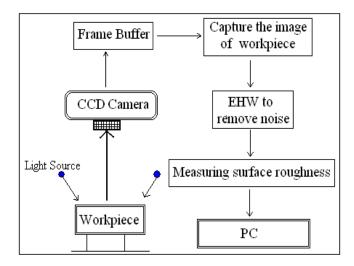


Fig.6 (b) Methodology in the proposed computer vision system for measuring surface roughness

A. Algorithm for feature extraction

- (i) Carry out image enhancement of machined image
- (ii) Subject the enhanced image to a L-level discrete wavelet decomposition.
- (iii) At each level (i=1, 2, ... L), there are four sub-images. One approximation image and three components/images (LH, HL and HH or

Horizontal, Vertical and Diagonal components). Calculate the weighted standard deviations of three detailed images.

$$f = \left\{ \sigma_{1}^{H}, \sigma_{1}^{V}, \sigma_{1}^{D}, \frac{1}{2} \sigma_{2}^{H}, \frac{1}{2} \sigma_{2}^{V}, \frac{1}{2} \sigma_{2}^{D}, \dots, \frac{1}{2} \sigma_{L}^{D}, \frac{1}{2^{L-1}} \sigma_{L}^{D}, \frac{1}{2^{L-1}} \sigma_{L}^{D}, \frac{1}{2^{L-1}} \sigma_{L}^{D}, \right\} \dots (14)$$

where, σ_i^M = Standard deviation of the M detail image at ith
Level

Level
M=H(Horizontal)/V(Vertical)/D(Diagonal)

component

The standard deviation of each sub image at level i is weighted by the factor (1/2i-1),

(iv) Repeat steps 1-4 four times for original image and images at orientation 90°, 180°, and 270 ° (achieved by rotating original image).

The final feature set consists of 4*(3L) features.

B. Wavelet based Feature Extraction

Since, the wavelet coefficient are orthogonal, the original profile can be re-obtained after wavelet decomposition by simply adding the sub-scales signals as shown in Figure 7. Furthermore, using this simple summation technique the concepts of roughness, waviness and form can be preserved. This is reflected in Figure 8 where, an arbitrary decomposition of a surface texture is obtained by casting into three frequency components, representing the form, waviness and roughness, using Daubechies wavelet of order 20. A dimensional step can now be cleared. Indeed, the same kind of decomposition process can be performed using images instead of profiles, because surface roughness can be measured precisely using for instance optical surface measurement systems. The arbitrary decomposition into form waviness and roughness of surface textures obtained by casting, grinding and vertical milling respectively, using wavelet of order 20 is shown. The roughness average of each component (i.e. form, waviness and roughness) is also shows in order to illustrate the roughness scale. The measured area is of a few millimeters square. Hence, the wavelet tool allows the decomposition of surfaces into form, waviness and roughness components and can successfully replace standard filters that are commonly used in surface texture characterization and hence, give a solid theoretical base for the standardization of these filters.

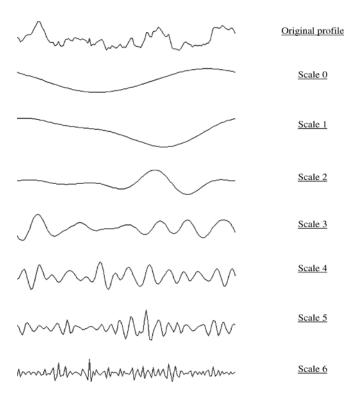


Fig. 7 Multiscale decomposition of a surface texture profile obtained by casting under seven different scales using the wavelet of order 20

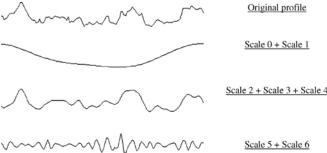


Fig. 8 Multiscale decomposition of a surface texture profile obtained by casting under three different components (form waviness and roughness) using the wavelet of order 20.

In Figure 9, the FNWT maxima indicate at each scale the location of a frequency component. Those features can also be quantified according to both the shape of the corresponding peak and its height. For an image, when using a multiresolution scheme for a dyadic standard decomposition of a function into sub-bands a filter bank with a power of two number of filters should be used. When using orthogonal wavelets like ones, one can easily simplify the problem by gathering the channels by scale in both directions. This process applied to a discrete wavelet is called the scaled DWT. The frequency normalization can then be performed based on these filters.

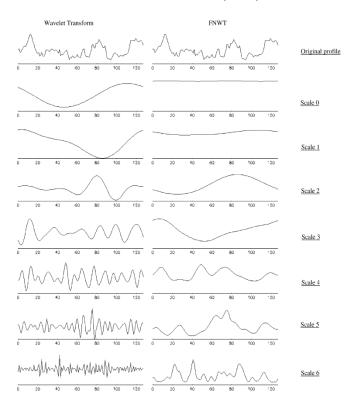


Fig. 9 Comparison of multiscale decomposition of a surface texture profile obtained by casting under seven different scales using the wavelet of order 20 and its frequency normalized equivalent.

C. ANN based surface roughness estimation

ANN with a variation of the classic back-propagation algorithm is employed to predict surface roughness. Compared with more conventional approaches, ANN demonstrates certain advantages that make them much more attractive. They have the ability to recognize patterns that are similar, but not identical, it can store information and generalize it. There is no need for explicit statement of the problem or for a problem-solving algorithm. Due to their massive parallelism, ANNs display increased computational power that can be used to deal with complex problems. Back-propagation neural network used for estimating the surface roughness of the machined surfaces with is a four layer network with six nodes in the input layer, six nodes in the first hidden layer, five nodes in the second hidden layer and one single node in the output layer.

Each layer is fully connected to the succeeding layer. The outputs of nodes of one layer are transmitted to nodes in another layer through links. The structure of an ANN is shown in Figure 10 where the Energy maps are fed as inputs into the trained neural network and the surface roughness parameter (R_t) is estimated . In the training phase, the desired value of the node in the output layer is the actual roughness value, R_t obtained by stylus method. The ANN adjusts the weights in all connecting links such that the mean square error, i.e. the averaged squared error

between the network output and the desired output is minimized. Training of ANN is stopped as soon as the specified number of epochs has reached and the values of weights corresponding to the minimum error are restored. Once trained, the ANN is then tested for different sets of input data. In the testing phase of the neural network, the predicted roughness, R_t is the value of the node in the output layer.

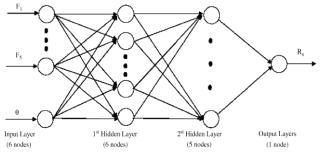
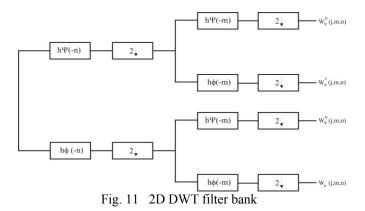


Fig. 10 The System architecture of ANN used for predicting Ra for surfaces

VI. RESULTS AND DISCUSSION

Case 1: Feature Extraction using Scaled DWT

The blocks contain time reversed scaling and wavelet vectors. The $h_{\phi}(-n)$ and $h_{\Psi}(-m)$ are low pass and high pass decomposition filters. Blocks are containing a down arrow and represent down sampling extracting every other point from a sequence of points. Each pass through the filter bank in Figure 11 decomposes the input signal into four lower resolutions (or lower scale) components. The W_{ϕ} coefficients are created by two low pass (h_{ϕ} based) filters and are thus called the approximation coefficients and $\{W_{\phi}^{i}\}$ for i = H, V, D are the horizontal, vertical and diagonal detail coefficients.



Mathematically, the series of filtering and down sampling operations are used to compute the DWT coefficients W_{ϕ} (j,m,n) and $\{W_{\phi}^{i}(j,m,n) \text{ for } i=H,V,D\}$ at scale j.

$$W_{\psi}^{H}(j,m,n) = h_{\psi}(-m) * \left[h_{\psi}(-n) * W_{\psi}(j+1,m,n)/n = 2k, k \ge 0 \right]$$

$$(m=2k), k \ge 0 \qquad ... (14)$$

$$W_{\psi}^{V}(j,m,n) = h_{\psi}(-m) * \left[h_{\psi}(-n) * W_{\psi}(j+1,m,n)/n = 2k, k \ge 0 \right]$$

$$(m=2k), k \ge 0 \qquad ... (15)$$

$$W_{\psi}^{D}(j,m,n) = h_{\psi}(-m) * \left[h_{\psi}(-n) * W_{\psi}(j+1,m,n)/n = 2k, k \ge 0 \right]$$

$$(m=2k), k \ge 0 \qquad ... (16)$$

$$W_{\psi}(j,m,n) = h_{\psi}(-m) * \left[h_{\psi}(-n) * W_{\psi}(j+1,m,n)/n = 2k, k \ge 0 \right]$$

$$/(m=2k), k \ge 0 \qquad ... (17)$$

In Figure 12, f(x,y) is the highest resolution representation of the image being transformed. It serves as the input for the first iteration and for the succeeding iterations; the approximation coefficients $W_{\phi}(j, m, n)$ are given as input to the filter bank, to obtain the next set of wavelet coefficients.

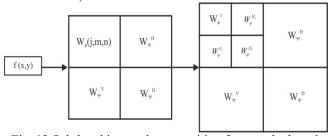


Fig. 12 Sub-band image decomposition for wavelet based feature extraction

Thus the energy for each subband is calculated up to 4 levels of decomposition and the image features Et, Eh, Ev and Ed are obtained from the energy map which is determined using tree-structured wavelet transform for each image. Few Sample enhanced machine images [Figure 13(a) to 18(a)] are applied with DWT and the respective transform outcomes are shown in [Figure 13(b) to 18(b)] along with the energy details in Table 1.

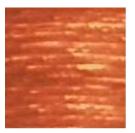


Figure 13(a)

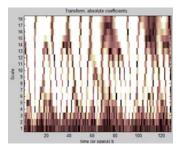


Figure 13(b) Transform absolute coefficient

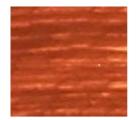


Figure 14(a)

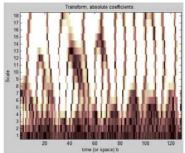


Figure 14(b) Transform absolute coefficient

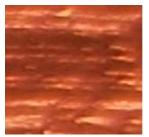


Figure 15(a).

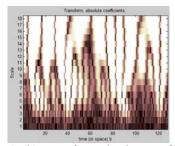


Figure 15(b) Transform absolute coefficient

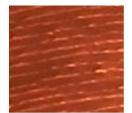


Figure 16(a)

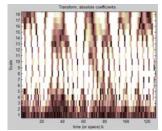


Figure 16(b) Transform absolute coefficient

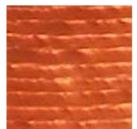


Figure 17(a).

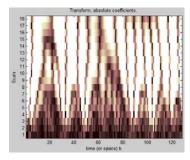


Figure 17(b) Transform absolute coefficient

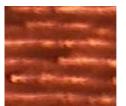


Figure 18(a)

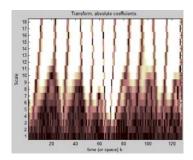


Figure 18(b) Transform absolute coefficient

Table 1 Energy maps obtained from DWT

Fig.	Ea	Sy maps c	Edetail		
Name	Values		Luctan	varues	
Name	values	Ε4	E1.	Е	E.J
		Et	Eh	Ev	Ed
13(a)	99.0286	0.0396	0.1714	0.4482	0.3122
14(a)	99.0088	0.0204	0.1183	0.3862	0.4663
15(a)	97.7414	0.0272	0.2686	0.6214	1.3414
16(a)	98.5546	0.0393	0.3785	0.7579	0.2697
17(a)	98.5984	0.0606	0.2933	0.5217	0.5260
. (-/)					
18(a)	96.8104	0.0282	0.1311	0.5938	2.4364
10(4)	70.0101	0.0202	0.1511	0.5750	2.1301
	1				

Where Et is Energy total, Eh is Energy horizontal, Ev is Energy Vertical and Ed is Energy diagonal. Ea is Energy Approximation.

Case 2: Estimation of R_t using ANN (a) For Milled surfaces

Two types of feature extraction and surface roughness estimation using ANN is performed in this work. The first one extracts the features using FT and the second uses the WT. In FT approach the key input features collected for training the network consist of

- (i) average grey scale value (Ga)
- (ii) major peak frequency (F1) and
- (iii) Principal component magnitude squared value (F2).

The WT based feature extraction is already discussed in case (i) of section VI. In the training phase (for both FT and WT) the desired value of the node in the output layer is the surface roughness R_t obtained using the stylus method. The surface roughness R_t from ANN along with the stylus measurement values for the milled samples after image enhancement with FT (WT) extracted features is given in Table 2 (Table 3).

Table 2 ANN estimated R_t for milling parameters (Image enhancement done and image features extracted using FT)

T4		ining con	ditions	Feature	s of imag	e texture	Stylus	Vision	
Test No.	V (m/s)	F (m/rev)	D (mm)	Ga	\mathbf{F}_1	F ₂	R _t (µm)	R _t (µm)	Error
1	250	40	1.6	0.5332	0.5552	0.6872	0.3016	0.31	0.0084
2	500	40	1.6	0.6605	0.5846	0.8016	0.1382	0.31	0.1718
3	500	80	1.6	0.5834	0.6537	0.7892	0.2823	0.47	0.1877
4	1000	40	0.4	0.5686	0.4686	0.8589	0.0877	0.20	0.1123
5	1000	80	0.8	0.5626	0.4342	0.9080	0.0567	0.17	0.1133
6	1000	160	0.4	0.5844	0.4326	0.8979	0.0528	0.17	0.1172
7	1000	160	0.8	0.6898	0.5864	0.9158	0.0260	0.13	0.104
8	2000	160	0.4	0.8307	0.8248	0.9443	0.0281	0.17	0.1419
9	355	80	1	0.4782	0.2627	0.6878	0.1217	0.07	0.0517
10	500	80	1	0.4858	0.2524	0.7814	0.0606	0.08	0.0194
11	710	500	1	0.8218	0.8119	0.2649	0.8486	0.83	0.0186
12	1400	500	1	0.9804	0.9642	0.6850	0.4649	0.37	0.0949

Table 3 ANN estimated R_t for milling parameters (Image enhancement done and image features extracted using WT)

		ining con	ditions	featu	res of i	mage te	xture			
Test No.	V (m/s)	F (m/rev)	D (mm)	Et	E _h	E _v	\mathbf{E}_{d}	Stylus R _t (µm)	Vision R _t (µm)	Error
1	250	40	1.6	0.9673	0.9085	0.7011	0.9644	0.3016	0.28	0.0216
2	500	40	1.6	0.9800	0.8131	0.8597	0.3587	0.1382	0.13	0.0082
3	500	80	1.6	0.9767	0.7354	0.5408	0.6609	0.2823	0.27	0.0123
4	1000	40	0.4	0.9870	0.8531	0.5585	0.2162	0.0877	0.07	0.0177
5	1000	80	0.8	0.9918	0.5923	0.3659	0.1265	0.0567	0.04	0.0167
6	1000	160	0.4	0.9928	0.5700	0.3303	0.1038	0.0528	0.04	0.0128
7	1000	160	0.8	0.9953	0.4977	0.2527	0.0436	0.0260	0.05	0.024
8	2000	160	0.4	0.9940	0.3938	0.3130	0.0743	0.0281	0.05	0.0219
9	355	80	1	0.9969	0.1338	0.1555	0.0432	0.1217	0.11	0.0117
10	500	80	1	0.9968	0.1438	0.1688	0.0404	0.0606	0.06	0.0006
11	710	500	1	0.9955	0.1646	0.1891	0.0842	0.8486	0.83	0.0186
12	1400	500	1	0.9908	0.2031	0.2416	0.2497	0.4649	0.45	0.0149

The results obtained are validated by plotting the correlation graph between stylus measured (conventional method) $R_{\rm r}$ and vision measured (proposed) $R_{\rm t}$ for both the FT and WT techniques for milled components is shown in Figure 15.

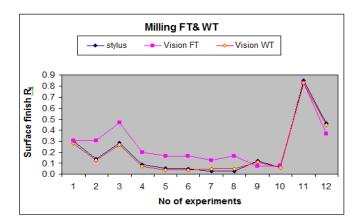


Figure 15 Comparison between predicted roughness values using vision approach and stylus approach for FT features and WT features (milling)

(b) For grinding operations

The R_t value predicted using the trained ANN and that measured using the stylus approach for the grinding process after image enhancement with features extracted using FT (WT) is given in Table 4 (Table 5).

Table 4 ANN estimated R_t for grinding parameters (Image enhancement done and image features extracted using FT)

Test	Mac	hining cond	itions	Roughne	ess features texture	of image	Stylus	Vision	Error
No.	V (m/s)	F (m/rev)	D (mm)	Ga	F ₁	F ₂	R _t (µm)	R _t (µm)	
1	23.55	10	15	0.7567	0.6924	0.9215	0.3536	0.29	0.0636
2	23.55	10	25	0.9697	0.8559	0.3064	0.4731	0.52	0.0469
3	23.55	15	15	0.7135	0.9985	0.1160	0.4388	0.41	0.0288
4	26.17	10	25	0.6239	0.4969	0.7461	0.2942	0.34	0.0458
5	26.17	15	25	0.5467	0.5704	0.2564	0.2196	0.28	0.0604
6	32.71	10	5	0.4251	0.2619	0.9999	0.3154	0.38	0.0646
7	36.63	5	20	0.7280	0.5521	0.2830	0.1570	0.15	0.007
8	36.63	10	10	0.3497	0.5164	0.2163	0.1368	0.21	0.0732
9	36.63	10	20	0.6588	0.5545	0.1020	0.1521	0.18	0.0279
10	39.25	5	20	0.6268	0.5296	0.0873	0.1573	0.13	0.0273
11	39.25	10	10	0.3858	0.4841	0.1208	0.1719	0.16	0.0119
12	39.25	15	25	0.6685	0.6458	0.2153	0.1825	0.15	0.0325

Table 5 ANN estimated R_t for grinding parameters (Image enhancement done and image features extracted using WT)

	Machi	ning condi	tions	Feat	tures of in	nage textu	ire	Stylus	Vision	Error
Test No.	V (m/sec)	F (m/sec)	D (mm)	Et	Eh	E _v	E _d	R _t (μm)	R _t (μm)	
1	23.55	10	15	0.9714	0.1152	0.7581	0.535	0.3536	0.36	0.0064
2	23.55	10	25	0.9596	0.1065	0.7584	0.835	0.4731	0.49	0.0169
3	23.55	15	15	0.9422	0.0675	0.4006	0.289	0.4388	0.45	0.0112
4	26.17	10	25	0.9799	0.0983	0.9592	0.320	0.2942	0.27	0.0242
5	26.17	15	25	0.9714	0.0896	0.6078	0.594	0.2196	0.19	0.0296
6	32.71	10	5	0.9865	0.1038	0.6269	0.194	0.3154	0.30	0.0154
7	36.63	5	20	0.9965	0.0531	0.2090	0.056	0.1570	0.17	0.013
8	36.63	10	10	0.9958	0.0474	0.3479	0.066	0.1368	0.18	0.0432
9	36.63	10	20	0.9960	0.0566	0.2660	0.068	0.1521	0.16	0.0079
10	39.25	5	20	0.9981	0.0156	0.1673	0.037	0.1573	0.16	0.0027
11	39.25	10	10	0.9992	0.0152	0.1667	0.034	0.1719	0.16	0.0119
12	39.25	15	25	0.9918	0.0195	0.1904	0.068	0.1825	0.19	0.007:

The results obtained are validated by plotting the correlation graph between stylus measured (conventional method) $R_{\rm r}$ and vision measured (proposed) $R_{\rm t}$ for both the FT and WT techniques for grinding components is shown in Figure 16.

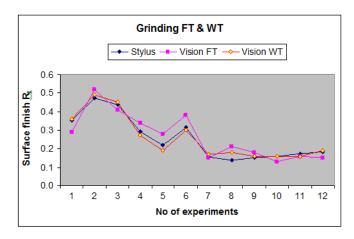


Figure 16 Comparison between predicted roughness values using vision approach and stylus approach for FT features and WT features (grinding).

VII. CONCLUSION AND FUTURE ENHANCEMENT.

The developed model is tested online on images of specimens grabbed by computer vision systems with linearly decreasing intensity. The features of the grabbed enhanced image (to remove noise present in the captured image) are extracted using two different schemes, one using Fourier transform (FT) and the other using wavelet

decomposition. The FT method is used to extract the features of image texture, namely, the major peak frequency F_1 , and the principal component magnitude squared value F_2 . Using the wavelet (Db4) multi resolution decomposition algorithm, the energy details of the sub band images, namely, energy total (E_t), energy horizontal (E_h), energy vertical (E_v) and energy diagonal (E_d) are extracted. These extracted features of the enhanced image are given as input to a trained neural network (back propagation network) and the surface roughness parameter R_t is estimated. From the obtained results, it is concluded that the wavelet based image feature extraction of the enhanced images gives better correlation between vision R_t and the stylus R_t both for milled and grinding surfaces.

Future direction of research shall focus on implementing the proposed algorithms using high speed hardware units thus making the present work ideally for high speed real-time machine vision applications.

VIII REFERENCES

- [1] G.A. Al-Kindi, R.M. Baul, K.F. Gill, An application of machine vision in the automated inspection of engineering surfaces, International Journal of Production Research 30 (2) (1992) 241–253
- [2] M.B. Kiran, B. Ramamoorthy, B. Radhakrishnan, Evaluation of surface roughness by vision system, International Journal of Machine Tools & Manufacture 38 (5–6) (1998) 685–690.
- [3] Du-Ming Tsai, Jeng-Jong Chen, Jeng-Fung Chen, A vision system for surface roughness assessment using neural networks, International Journal of Advanced Manufacturing Technology 14 (6) (1998) 412–422.
- [4] M.Y. Rafiq, G. Bugmann, D.J. Easterbrook, Neural network design for engineering applications, Computers and Structures 79 (17) (2001) 1541–1552.
- [5] K. Venkata Ramana, B. Ramamoorthy, Statistical methods to compare the texture features of machine surfaces, Pattern Recognition 29 (9) (1996) 1447– 1459.
- [6] P.G. Benardos, G.C. Vosniakos, Prediction of surface roughness in CNC face milling using neural networks and Taguchi's design of experiments, Robotics and Computer Integrated Manufacturing 18 (5–6) (2002) 343–354.
- [7] Shengyu Fu, B. Muralikrishnan, J. Raja, "Engineering Surface Analysis with different wavelet bases", Trans. of ASME, vol 125, Nov. 2003, pp. 844-852.
- [8] H.T. Hingle and J.H. Rakels, The practical application of diffraction techniques to assess surface finish of diamond turned parts, Ann. CARP, 32(1)(1983)499-501.
- [9] B. Josso, D.R. Burton, M.J. Lalor, Wavelet strategy for surface roughness analysis and characterisation, Comput. Methods Applications. Mech. Eng. 191 (8– 10) (2001) 829–842.

- [10] Daubechies, The wavelet transform, time-frequency localization and signal analysis, IEEE Trans. Inform. Theory 36 (1990) 961–1005.
- [11] Xiaodong Gu, Daoheng Yu, Liming Zhang, Image shadow removal using pulse coupled neural network, IEEE Transactions on Neural Networks 16 (3) (2005) 692–698.
- [12] X.Q. Jiang, L. Blunt, K.J. Stout, Three-dimensional surface characterization for orthopaedic joint prostheses, proceedings of institution of mechanical engineers. Part H, J. Eng. Med. 213 (1) (1999) 49–68.
- [13] Grzesik W., Rech J., Wanat T.: Comparative study of the surface roughness produced in various hard machining processes. 3rd International Congress of Precision Machining, Vienna, Austria, 2005, pp. 119-124.
- [14] Josso B., Burton D., Lalor M.: Frequency normalized wavelet transform for surface roughness analysis and characterization. Wear, Vol. 252, 2002, pp. 491-500.
- [15] S.S. Liu, M.E. Jernigan, Texture analysis and discrimination in additive noise, Computer Vision, Graphics and Image Processing 49 (1) (1990) 52–67.
- [16] Zawada-Tomkiewicz A., Storch B.: Introduction of the wavelet analysis of a machined surface profile. Advances in Manufacturing Science and Technology, Vol. 28, No. 2, 2004, pp. 91-100.
- [17] S.-H. Lee, H. Zahouani, R. Caterini, T.G. Mathia, Morphological characterization of engineered surfaces by wavelet transform, in: Proceedings of the 7th International Conference on Metrology and Properties of Engineering Surfaces, Götebarg, Sweden, 1997, pp. 182–190.

IX AUTHORS PROFILE

- 1. Mr. T.K.Thivakaran is presently a research scholar in MS university, Thirunelveli in the faculty of Computer Science and Engineering. He is working as Assistant Professor in the faculty of Information Technology, Sri Venkateswara college of Engineering, Chennai. His area of research includes Image Processing, Cryptography and Network Security.
- 2. Dr.RM.Chandrasekaran is currently working as a Professor at the Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar, Tamilnadu, India. From 1999 to 2001 he worked as a software consultant in Etiam, Inc, California, USA. He received his Ph.D degree in 2006 from Annamalai University, Chidambaram. He has conducted workshops and conferences in the area of Multimedia, Business Intelligence, Analysis of Algorithms and Data Mining. He has presented and published more than 32 papers in conferences and journals and is the co-author of the book Numerical Methods with C++ Program(PHI,2005). His research interests include Data Mining, Algorithms, Image processing and Mobile Computing. He is life member of the Computer Society of India, Indian Society for Technical Education, Institute of Engineers etc.

An In-depth Study on Requirement Engineering

Mohammad Shabbir Hasan¹, Abdullah Al Mahmood², Farin Rahman³, Sk. Md. Nahid Hasan⁴

Panacea Research Lab, Dhaka, Bangladesh

Abstract— Software development includes Requirement Engineering (RE) which is comprised of discovering stakeholders and their need, proper documentation, communication and subsequent implementation. It can be viewed as the most crucial phase as the success of the software depends largely on it. Requirement Engineering is receiving increasing attention of the researchers and also people associated with software development. This paper investigates RE activities in detail followed by some current challenges and also proposes some suggestions to overcome these challenging issues.

Keywords- Software Requirement, Requirement Engineering, Requirement Elicitation, Requirement Management.

I. INTRODUCTION

Success of a software system is measured by evaluating how it meets the purpose it was intended and in broad sense, Requirement Engineering (RE) is the process of discovering that purpose through identifying stake holders and their needs, refinement of the gathered information, modeling, specification and then subsequent implementation. The system requirements and role allocated to software—initially established by the system engineer—are refined in detail. Models of the required data, information and control flow, and operational behavior are created. Alternative solutions are analyzed and a complete analysis model is created [1]. Donald Reifer [2] describes the software requirement engineering process in the following way:

"Requirements engineering is the systematic use of proven principles, techniques, languages, and tools for the cost effective analysis, documentation, and on-going evolution of user needs and the specification of the external behavior of a system to satisfy those user needs. Notice that like all engineering disciplines, requirements engineering is not conducted in a sporadic, random or otherwise haphazard fashion, but instead is the systematic use of proven approaches."

The Requirement Engineering process may face a number of inherent difficulties like: stakeholders may be numerous and distributed, their goals may be volatile and conflicting, depending on their perspectives of different working environment, goals may be implicit and difficult to articulate, and, inevitably, satisfaction of these goals may be

constrained by a variety of factors outside their control. Based on these characteristics Zave [3] defined RE as:

"Requirements engineering is the branch of software engineering concerned with the real-world goals for, functions of, and constraints on software systems. It is also concerned with the relationship of these factors to precise specifications of software behavior, and to their evolution over time and across software families."

This definition is an attractive one as it highlights the importance of "real world goals" which motivates the development of a software system by referring the specifications more precisely. It also refers to the reality of a changing world and the need to reuse partial specifications, as done by engineers in other branches of engineering. Brooks [4] defined RE as a key problem area in the development of complex, software-intensive systems:

"The hardest single part of building a software system is deciding what to build. ...No other part of the work so cripples the resulting system if done wrong. No other part is more difficult to rectify later."

RE can be characterized as a branch of *System Engineering* [5] as it has to encompass a systems level view to deliver some systems behavior to its stakeholders. Again, Zave defined RE as a branch of *Software Engineering* and software systems requirements engineering has received special consideration possibly due to the abstract and invisible nature of software with the vast range and variety of problems that admit to software solutions.

Whether viewed at the systems level or the software level, RE is a multi-disciplinary, human-centered process and the use of the term *Engineering* in RE serves as a reminder that RE is an important part of an engineering process and represents a series of engineering decisions that lead from recognition of a problem to be solved to a detailed specification of that problem followed by anchoring development activities, so that the appropriateness and cost-effectiveness of the solution can then be analyzed.

The structure of the paper is approximately as follows. In sections 2 and 3 we discuss the foundation and some ground works needed for requirement engineering respectively.

Section 4 discusses about some key issues followed by the core activities of RE in section 5. Then, in section 6, challenges of RE activities are analyzed with some suggestions. We finish by drawing some broad and necessarily speculative and personal conclusions about the future of requirement engineering.

II. FOUNDATION

RE activities play a vital role in software and systems engineering, and there are also many disciplines upon which it draws.

In software development context, an important role is played by Computer Science as theoretical computer science provide the framework to assess the feasibility of requirements while the means of developing software solution are provided by practical computer science. Here *logic* provides a vehicle for analyzing software behavior which is acquiescent to formal reasoning to make the reasoning steps more explicit. Different logics may be used to express different aspects of a required system. For example, *temporal logic* can be used to describe timing information, *deontic logic* to describe permissions and obligations, and *linear logic* to describe resources and their use. A further advantage of specification languages grounded in logic is that they are potentially amenable to automated reasoning and analysis [6].

In the systems engineering context, an understanding and application of systems theory and practice is also relevant to RE [5]. This includes work on characterizing systems, identifying their boundaries and managing their development life cycle [7-8]. RE also encompasses work on systems analysis, traditionally found in the information systems world [9].

Usually RE activities take place in a human activity system, and people are the problem owner. Therefore, RE needs to be sensitive to how people perceive and understand the newly implemented computer-based system, how they interact, and how the environment of the workplace is affected by their actions. RE draws on the cognitive and social sciences to provide both theoretical grounding and practical techniques for eliciting and modeling requirements [6]. Cognitive psychology provides an understanding of the difficulties people may have in describing their needs [10]. Anthropology provides a methodological approach to observing human activities that helps to develop a richer understanding of how computer systems may help or hinder those activities [11]. Sociology provides an understanding of the political and cultural changes caused by computerization. Introduction of a new computer system changes the nature of the work carried out within an organization, may affect the structure and communication paths within that organization, and may even change the original needs that it was built to satisfy [12]. Linguistics is important because RE is largely about communication. Linguistic analyses have changed the way in which the English language is used in specifications, for instance to avoid ambiguity and to improve understandability. Tools from linguistics can also be used in

requirements elicitation, for instance to analyze communication patterns within an organization [13]. *Philosophical* elements also have an important effect on RE as it is concerned with interpreting and understanding of terminology, concepts, viewpoints and goals from stakeholder's perspective. Such issues become important while requirement validation, especially where stakeholders may have divergent goals and incompatible belief systems. They also become important in selecting a requirement modeling technique, because the choice of technique affects the set of phenomena that can be modeled, and may even restrict what a requirements engineer is capable of observing [6].

III. GROUNDWORK

In the software development process, RE is considered as a front-end activity. Although it is usually the case that requirements change during development and evolve after a system has been in operation for some time, RE plays an important role in the management of changes in software development. Nevertheless, the bulk of the effort of RE does occur early in the lifetime of a project, motivated by the evidence that requirements errors, such as misunderstood or omitted requirements, are more expensive to fix later in project lifecycles [14-15].

Before a project can be started, some preparation is needed which are categorized as *context and groundwork* by Finkelstein [16]. This groundwork includes some assessment of project's feasibility and associated risks needs to be undertaken, and RE plays a crucial role in making such an assessment. Although it is often possible to estimate project costs, schedules and technical feasibility from precise specifications of requirements, risk should be reevaluated regularly throughout the development lifetime of a system [17], since changes in the environment can change the associated development risks.

Again, RE activities are performed in a variety of contexts, including market-driven product development and development for a specific customer with the eventual intention of developing a broader market. The type of product also affects the choice of method: RE for information systems is very different from RE for embedded control systems, which is different again from RE for generic services such as networking and operating systems [6]. So groundwork is essential for the identification of a suitable process and also for the selection of suitable methods and techniques for the various RE activities.

IV. KEY ISSUES IN REQUIREMENT ENGINEERING

Software development organizations should keep in mind the following issues when they consider how to improve the requirements and communication for their projects:

 If teams don't get requirements right, it doesn't matter how well they execute the rest of the project: The goal of every software development project is to build a product that provides value to customers. Effective requirements definition enables teams to determine the mix of product capabilities and characteristics that will best deliver this customer value. An understanding evolves over time as stakeholders provide feedback on the early work and refine their expectations and needs. Adequately exploring and crafting requirements into a set of product features and attributes helps to ensure customer needs are being met throughout the project lifecycle [18].

- Requirements definition is a discovery and invention process, not just a collection process: Teams often talk about "gathering requirements." This phrase suggests that requirements are just lying around waiting to be picked like flowers or to be plucked out of users' brains by an analyst. In reality, requirements definition is an exploratory activity, and requirements elicitation is a more accurate description than requirements gathering. Elicitation includes some discovery and some invention, as well as recording those bits of requirements information that various stakeholders present to an analyst. Elicitation demands iteration. Constant feedback and validation from stakeholders keeps communication Participants in a requirements elicitation discussion will not think of everything they will need up front, and their thinking will change as a project progresses. Teams that prepare to iterate most often elicit the most accurate requirements [18].
- Customer involvement is the most critical contributor to software quality: Various studies confirm that inadequate customer involvement is a leading cause of failure of software projects. The development team will get the customer input it needs eventually – even if it is after a project ships. However, it is much cheaper – and much less painful – to get customer input earlier, rather than after product release. Customer involvement requires more than a workshop or two early in the project. It involves input from customers early and often in the requirements process. Ongoing engagement by suitably empowered enthusiastic stakeholders is a critical success factor for software development [18].
- Change happens; managing change is critical: It is inevitable that requirements will change as business needs evolve, new users or markets are identified, business rules and government regulations are revised and operating environments change. The objective of a change control process is not to inhibit change. Rather, the objective is to manage change to ensure that the project incorporates the right changes for the right reasons. Teams that anticipate and accommodate changes minimize disruption and cost to the project and its

- stakeholders. Further teams that can force as much change at the beginning of a project will have less change to manage over time [18].
- Teams are never going to have perfect requirements: Requirements are never finished or complete. There is no way to know for certain that teams have not overlooked some requirement, and there will always be some requirements that are not in the specification. It is also folly to think teams can freeze the requirements and allow no changes after some initial elicitation activities. Rather than declaring requirements "done" at some point, effective teams define a baseline then follow a sensible change control process to modify requirements once a baseline is established [18].

V. CORE ACTIVITIES OF REQUIREMENT ENGINEERING

Good RE activities can accelerate software development. The process of defining business requirements aligns the stakeholders with shared vision, goals and expectations. Involvement of substantial user in establishing and managing changes to agree upon requirements increases the accuracy of requirements, ensuring that the functionality of the developed system will enable users to perform their essential business tasks.

In a broad sense, Requirement Engineering encompasses the two major sub domains of requirement definition and requirement management.

Requirement Definition is the collaborative process of collecting, documenting and validating a set of requirements that constitute an agreement among key project stakeholders. This phase can be further subdivided into the critical process areas of elicitation, analysis, specification, agreeing and validating requirements.

From a pragmatic perspective, requirement definition strives for requirements which are validated by user and clear enough to the software development team to proceed with design, construction and testing at an acceptable level of risk. It is natural that, risk leads to the threat of doing expensive and unnecessary rework.

Requirement Management involves working with a defined set of requirements throughout the development process of the system and its operational life. It also allows managing changes to that set of requirements throughout the project lifecycle. This phase includes selecting changes to be incorporated within a particular release and ensuring effective implementation of changes with no adverse impact on schedule, scope or quality of the developed system.

An effective requirement definition and management solution creates an accurate and complete set of system requirements, while helping organizations to improve communications is an effort to better align it with business needs and objectives. The following sub sections discuss the core activities of RE.

A. Requirement Elicitation

Requirements elicitation is recognized as one of the most critical, knowledge-intensive activities of software development [19]; poor execution of elicitation will lead a complete failure of the project. Since project failures are so rampant [20], it is quite likely that improving how the industry performs elicitation could have a dramatic effect on the success record of the industry [21]. Information collected during this requirement elicitation phase are interpreted, analyzed, modeled and validated so that a requirement engineer can feel confident that a complete enough set of requirements of a system have been collected. Therefore, requirements elicitation is closely related to other RE activities - to a great extent, selection of techniques for the next phases largely depends on how requirements were elicited. Steps in requirement elicitation phase are described below:

- Defining System Boundaries: One of the key objectives of requirement elicitation is to define system boundaries i.e., to find out the problems that need to be solved. This definition is necessary to discover where the final delivered system will fit into the current operational environment. Without a clearly defined system boundary, the project is issuing an open invitation to scope creep. Prior to eliciting requirements, teams should have a clear understanding of system boundaries as it affects all subsequent elicitation efforts.
- Defining Goals: Goals denote the objectives that a system should meet. Eliciting high level goals in the early stage of the development process is crucial. However, goal-oriented requirements elicitation [23] is an activity that continues as development proceeds, as high-level goals (such as business goals) are refined into lower level goals (such as technical goals that are eventually operational in a system) [6]. Eliciting goals emphasizes on defining the problem domain and also the needs of the stakeholders, rather than on possible solutions to those problems.
- *Identifying* Stakeholders: Stakeholders mean individuals or organizations that stand to gain or lose from the success or failure of a system. Stakeholders include customers or clients (who pay for the system), developers (who design, construct and maintain the system), and users (who interact with the system to get their work done) [6]. It is essential to gain commitment from key stakeholders for their participation throughout the requirements definition and Customer engagement is necessary during requirements management, as well. [18]. Stakeholders' view is required in making change in decisions followed by assessing the impact of proposed changes and adjusting requirement priorities. Users themselves are not

- homogeneous, and hence part of the elicitation process is to identify the needs of different user classes, such as novice users, expert users, occasional users, disabled users, and so on [22]. So initially every software project should identify its key requirement decision makers as well as the decision-making process to ensure that the right people can make important and timely decisions.
- Select Elicitation Technique: The choice of elicitation technique depends on the time and resources available to the requirements engineer, and of course, the kind of information that needs to be elicited [6]. It also depends on the extent of stakeholder involvement and how much access the analyst has to the stakeholders. To interact with stakeholders, requirement engineer can techniques like: workshops, questionnaires, and interviews. Now a day, it is a common practice that teams want to interact with stakeholders in lightweight, dynamic, frequent way which is completely focused on the problem and thus collaboration has taken priority documentation review. Techniques that leverage prototypes, mockups and screenshots are becoming the norm. However, for the sake of a dynamic development process, teams should be trained and proficient in a variety of elicitation techniques.
- Exploring user scenarios and simulations: It is often the case that users find it difficult to articulate their requirements. Rather they find it more convenient to visually describe their business tasks, patterns and expected interaction product functionality than to define all these textually. So a requirements engineer can resort to eliciting information about the tasks currently performed by the users and those that they might want to perform [24]. After that, these tasks can be represented in use cases which can be used to describe the outwardly visible requirements of systems [25]. More specifically, the requirements engineer may choose a particular path through a use case, a scenario, in order to better understand some aspect of using a system [26]. Acceleration in visual techniques for requirements definition and the ability to tie artifacts to more traditional requirements is changing the way of interactions forming between business and development organizations.

A-1. Elicitation Techniques:

Researchers found various classes for requirement elicitation techniques all of which are apposite in different scenarios. Some of these are briefly discussed below:

Traditional Techniques: These techniques are useful for gathering generic data. Questionnaires and surveys, interviews, and analysis of existing documentation (ex: organizational charts, process

- models or standards, and manuals of existing systems) belong to this group.
- Elicitation through Groups: These techniques are useful for a richer understanding of need through exploiting team dynamics with an aim to foster stakeholders' need. This class includes techniques like brainstorming and focus groups, as well as RAD/JAD workshops (using consensus-building workshops with an unbiased facilitator) [27].
- *Prototyping:* These techniques are useful in such cases where there remains a great deal of uncertainty about the requirements, or where early feedback from stakeholders is needed [28]. For better requirement elicitation, prototyping can also be readily combined with other techniques, for instance by using a prototype to provoke discussion in a group elicitation technique or as the basis for other traditional techniques.
- Model Driven Techniques: Techniques of this class is helpful in the cases where a specific model of the type of information to be gathered is already provided and this model is used to drive the elicitation process. Goal-based methods, such as KAOS [29] and I* [30], and scenario-based methods such as CREWS [31] belong to this class.
- Cognitive Techniques: This group includes a series of techniques originally developed for knowledge acquisition for knowledge-based systems [32]. Such techniques include protocol analysis (in which an expert thinks aloud while performing a task, to provide the observer with insights into the cognitive processes used to perform the task), laddering (using probes to elicit structure and content of stakeholder knowledge), card sorting (asking stakeholders to sort cards in groups, each of which has name of some domain entity), repertory grids (constructing an attribute matrix for entities, by asking stakeholders for attributes applicable to entities and values for cells in each entity) [6].
- Contextual Techniques: This class is an alternative to both traditional and cognitive techniques [33]. These include the use of ethnographic techniques like participant observation, ethnomethodogy and conversation analysis, which result fine grained analysis to identify patterns in conversation and interaction [34].

Though there is a fundamental methodological disagreement between the proponents of contextual techniques on the one hand, and the traditional and cognitive techniques on the other, recent work has focused on the question of whether integration is possible [34-35].

B. Requirement Analysis

Requirements Analysis is a fundamental activity in RE that bridges the gap between system level requirements

engineering and software design. It allows the requirement engineer to refine the software allocation and build models of the data, functional, and behavioral domains that will be exploited by software. It is also beneficial for the software designer as it provides a representation of information, function, and behavior that can be translated to data, architectural, interface, and component-level designs. Finally, requirements specification after successful requirement analysis provides the developer as well as the customer with the means to assess quality once software is built. The subsequent steps should be followed during requirement analysis:

- Creation of visual scenarios: The natural language requirements found in most specifications are mostly text based, full of ambiguities, redundancies and gaps. Most of the cases, it is desirable to represent requirements in multiple ways to give readers a richer, more holistic understanding. Visual scenarios present requirements information from a business viewpoint in graphical diagrams which allow reviewers to immediately spot missing requirements by examining flows, rather than uncovering missing requirements by reading a opaque textual specification. Teams may have better results using diagrams that communicate at a higher level of abstraction, so readers can get the big picture without getting mired in all of the details.
- Creation and Evaluation of Simulations: A simulation is an interactive software experience that captures the essential flow and level of detail requirements. Simulations opportunities for everyone to interact with some portion of the final system which is more tangible than written requirements specifications. A simulation may look and behave like a prototype, but the key difference is that a simulation is not created by the development team. It is created as part of the requirements definition phase and is typically done without requiring any development skills. Simulations can leverage existing business data, show flow of the data and dynamically capture feedback that can quickly be incorporated into the next revision. Typically, revisions can be augmented during the stakeholder review - leading to an "Is this what you had in mind?" high quality interaction [18].
- Requirement Prioritization: The ultimate goal of a
 software development team is to create a system
 that meets the stakeholders' demands. Since there
 are usually more requirements than can be
 implemented within the limited resource, decision
 makers must face the dilemma of selecting the
 right set of requirements for the intended system.
 In order to select the correct set of requirements,
 the decision makers must understand the relative

priorities of the requested requirements [36]. By selecting a subset of the requirements that are valuable for the customers, and can be implemented within budget, organizations can become more successful on the market. However, Prioritization should be a collaborative process that involves both a customer and a technical perspective to balance customer value against cost and technical risk.

Analysis techniques that have been investigated in RE include requirements animation [37], automated reasoning (e.g., analogical and case-based reasoning [38] and knowledge based critiquing [39]), consistency checking (e.g., model checking [40]), and a variety of techniques for validation and verification.

C. Requirement Specification

Software Requirement Specification (SRS) is basically an organization's understanding (in writing) of a customer or potential client's system requirements and dependencies at a particular point in time (usually) prior to any actual design or development work. It's a two-way insurance policy that assures that both the client and the organization understand the other's requirements from that perspective at a given point in time. The SRS document itself states in precise and explicit language those functions and capabilities a software system must provide, as well as states any required constraints by which the system must abide. The SRS also officiates as a blueprint for completing a project with as little cost growth as possible. The SRS is often referred to as the "parent" document because all subsequent project management documents, such as design specifications. statements of work, software architecture specifications, testing and validation plans, and documentation plans, are related to it. It is important to note that an SRS contains functional (requirements that define a function of a software system or its component) and nonfunctional requirements (requirements that specify criteria that can be used to judge the operation of a system, rather than specific behaviors) only; it doesn't offer design suggestions, possible solutions to technology or business issues, or any other information other than what the development team understands the customer's system requirements to be. There are some standard for SRS defined by IEEE e.g. IEEE STD 830-1998 [41] and IEEE STD 1233-1998 [42].

C-1. Goals of SRS

A well-designed, well-written SRS accomplishes four major goals:

• It provides feedback to the customer. An SRS is the customer's assurance that the development organization understands the issues or problems to be solved and the software behavior necessary to address those problems. Therefore, the SRS should be written in natural language, in an unambiguous manner that may also include charts, tables, data flow diagrams, decision tables, and so on.

- It decomposes the problem into component parts.
 The simple act of writing down software requirements in a well-designed format organizes information, places borders around the problem, solidifies ideas, and helps break down the problem into its constituent parts in an orderly fashion.
- It serves as an input to the design specification. As mentioned previously, the SRS serves as the parent document to subsequent documents, such as the software design specification and statement of work. Therefore, the SRS must contain sufficient detail in the functional system requirements so that a design solution can be devised.

It serves as a product validation check. The SRS also acts as the parent document for testing and validation strategies that will be applied to the requirements for verification and validation.

C-2. What should the SRS address?

IEEE standards suggest that SRS should address the following basic issues:

- Functionality: What is the software supposed to do?
- External interfaces: How does the software interact with people, the system's hardware, other hardware, and other software?
- *Performance:* What is the speed, availability, response time, recovery time of various software functions, etc.?
- *Attributes:* What are the portability, correctness, maintainability, security, etc. considerations?
- Design constraints imposed on an implementation: Are there any required standards in effect, implementation language, policies for database integrity, resource limits, operating environment(s) etc.?

C-3. Characteristics of a good SRS

A good SRS should contain the following characteristics:

- *Complete:* SRS should define precisely all the golive situations that will be encountered and the system's capability to successfully address them. It should contain everything that is needed by the software designers to develop the software.
- Correct: Of course everyone associated with the intended system expects the specification to be correct. No one writes a specification that they know is incorrect. Someone may like to say it "Correct and Ever Correcting." The discipline is keeping the specification up to date when someone finds things that are not correct.
- Consistent: The SRS should be consistent within itself and consistent to its reference documents. If someone calls an input "Start and Stop" in one

- place, he/she shouldn't call it "Start/Stop" in another.
- Accurate: SRS should precisely define the system's capability in a real-world environment, as well as how it interfaces and interacts with it. This aspect of requirements is a noteworthy problem area for many SRSs.
- Modifiable: The logical, hierarchical structure of the SRS should facilitate any necessary modifications; grouping related issues together and separating them from unrelated issues makes the SRS easier to modify.
- Ranked: Individual requirements of an SRS should be hierarchically arranged according to stability, security, perceived ease/difficulty of implementation, or other parameter that helps in the design of that and subsequent documents.
- Unambiguous: SRS must contain requirements statements that can be interpreted in one way only. This is another area that creates significant problems for SRS development because of the use of natural language.
- Testable: SRS must be stated in such a manner that unambiguous assessment criteria (pass/fail or some quantitative measure) can be derived from the SRS itself.
- Traceable: Requirements traceability (RT) is another major factor that determines how easy it is to read, navigate, query and change requirements documentation [6]. SRS must be able to link each software functional requirement back to its origin, possibly a use case or business rule. Teams should embrace traceability information that connects functional requirements to associated design elements, code segments and tests to accelerate debugging and software maintenance. As RT lies at the heart of requirements management practice, providing RT in SRS is a means of achieving integrity and completeness of requirement documentation that has an important role to play in managing changes.
- Valid: A valid SRS is one in which all parties and project participants can understand, analyze, accept, or approve it. This is one of the main reasons SRSs are written using natural language.
- Verifiable: A verifiable SRS is consistent from one level of abstraction to another. Most attributes of a specification are subjective and a conclusive assessment of quality requires a technical review by domain experts. Using indicators of strength and weakness provide some evidence whether preferred attributes are present or not.

C-4. Benefits of a good SRS

The IEEE 830 standard defines the benefits of a good SRS [41]:

- Establish the basis for agreement between the customers and the suppliers on what the software product is to do: The complete description of the functions to be performed by the software specified in the SRS will assist the potential users to determine if the software specified meets their needs or how the software must be modified to meet their needs.
- Reduce the development effort: The preparation of the SRS forces the various concerned groups in the customer's organization to consider rigorously all of the requirements before design begins and reduces later redesign, recoding, and retesting. Careful review of the requirements in the SRS can reveal omissions, misunderstandings, and inconsistencies early in the development cycle when these problems are easier to correct.
- Provide a basis for estimating costs and schedules:
 The description of the system to be developed as given in the SRS is a realistic basis for estimating project costs and can be used to obtain approval for bids or price estimates.
- Provide a baseline for validation and verification:
 Teams can develop their validation and
 Verification plans much more productively from a
 good SRS. As a part of the development contract,
 the SRS provides a baseline against which
 compliance can be measured.
- Facilitate transfer: SRS makes it easier to transfer the software product to new users or new machines. Customers thus find it easier to transfer the software to other parts of their organization, and suppliers find it easier to transfer it to new customers.
- Serve as a basis for enhancement: Because the SRS discusses the system but not the project that developed it, the SRS serves as a basis for later enhancement of the developed system. The SRS may need to be altered, but it does provide a foundation for continued system evaluation.

D. Agreeing and Validating Requirements

It is an arduous job to maintain agreement with all stakeholders as they have divergent goals. Validation is the process of establishing that the requirements elicited and specified provide an accurate account of stakeholder requirements. This activity is essential for resolving conflicts between stakeholders.

Techniques such as inspection and formal analysis tend to concentrate on the consistency and completeness of the requirements. This approach is illustrated by SCR [43] which provides an automated checking of the formal model

for syntactic consistency and completeness. On the other hand, techniques like prototyping, specification animation, and the use of scenarios are geared towards testing a correspondence with the real world problem: have all the aspects of the problem that the stakeholders regard as important been covered?

In some cases, the methods and tools used by the requirement engineers dominate the way that they see and describe problems, which, in the extreme case, shifts the problem of validating requirements statements to a problem of convincing stakeholders that the chosen representation for requirements models is appropriate. This leads to the realization that observation is not value free, rather it is theory-driven, and is biased by the current paradigm. Jackson captures this perspective through his identification of problem frames [44]. If stakeholders do not agree with the choice of problem frame, it is unlikely that they will ever agree with any statement of the requirements. Ethnomethodologists attempt to avoid the problem altogether, by refusing to impose modeling constructs on the stakeholders [33]. By discarding traditional problem analysis tools, they seek to apply value-free observations of stakeholder activities, and therefore circumvent the requirements validation issue altogether.

Requirements negotiation attempts to resolve conflicts on goal among stakeholders without necessarily weakening satisfaction of each stakeholder. Early approaches to requirements negotiation focused on modeling each stakeholder's contribution separately rather than trying to fit their contributions into a single consistent model [45] and on the importance of establishing common ground [46]. Boehm introduced the win-win approach [47] in which the win conditions for each stakeholder are identified, and the software process is managed and measured to ensure that all the win conditions are satisfied, through negotiation among the stakeholders. These negotiation models concentrate on the identification of the most important goals of each participant, and ensure that these goals are met. This approach is also used in other RE techniques for promoting agreement, without necessarily making the goals explicit [48]. For example, in Quality Function Deployment (QFD) [49], matrices are constructed to compare functional requirements with one another and rate their importance rather explicitly identifying stakeholder goals.

After agreeing of the requirements, teams should begin "testing" as soon as they have requirements in hand. Deriving test cases from use cases and scenarios is a valuable way to find problems in the use cases themselves, in functional requirements derived from the use cases and in analysis models created from the requirements [18].

E. Requirement Management

Managing change is a fundamental activity in RE [50] as every successful software system evolves due to change in the environment it is operating as well as change in the

requirement of stakeholders. Requirement Management should go through the following steps:

- Manage Requirements Versions: As requirements evolve during the course of a project, it is important to track the different versions of requirements specification documents and even individual requirements [18]. Changes to requirements documentation need to be managed properly and minimally, this involves providing techniques and tools for configuration management and version control [51], and by exploiting traceability, the impact of changes in different parts of the documentation can be monitored and controlled correctly. Version tracking helps to ensure that all team members are working from the latest requirements baseline.
- Establishing A Change Control Process: Typically, changes to requirement specifications include adding or deleting requirements, and also fixing errors. Requirements are added in response to changing stakeholder needs that were missed in the initial analysis. Requirements are deleted usually in the circumstances to forestall cost and schedule overruns. Again there also remain inconsistencies in requirements which arise both as a result of mistakes and because of conflicts between requirements. In any case, managing inconsistency in requirements specifications as they evolve is a major challenge [52]. So once requirements have been base lined, proposed modifications in them should follow an established change control process which provides consistency in the way requirement changes are proposed, evaluated, approved or rejected, communicated to stakeholders and implemented in affected work products. Teams should have formal written change control processes in place before eliciting requirements [18].
- Perform Requirements Change Impact Analysis:
 Requirement Management not only includes process of managing documentation but also process of recognizing change through continued requirements elicitation, re-evaluation of risk, and evaluation of systems in their operational environment. Thus, each proposed change needs to be evaluated in terms of existing requirements and architecture so that the trade-off between the cost and benefit of making a change can be assessed.

Finally, the process of identifying core requirements in order to develop architectures that are (a) stable in the presence of change, and (b) flexible enough to be customized and adapted to changing requirements, is one of the key research issues in software engineering [53].

VI. REQUIREMENT ENGINEERING: CHALLENGES AND SUGGESTIONS

As mentioned earlier, Requirement Engineering is a core process as well as most crucial part in software development life cycle. Bugs in requirements may not be identified during development phase rather they remain concealed until system becomes operational and customer requirements are not met. Poor requirements cause not only modifications in requirement specifications but also require re-designing, re-implementing and retesting for entire software. Therefore, requirement engineers have to struggle and conquer uncountable numbers of challenges for developing effective and efficient software.

Anticipating requirement engineering challenges will grant opportunities for requirement engineers to enhance software success rate. There have been many investigations conducted to explore different challenges in various domains of requirement engineering. Some suggestions to overcome these challenges are discussed below:

• Elicitation: Reduce Rework by Capturing Better Information

The process of capturing requirements in requirements elicitation includes context, technical identifying key business and stakeholders, getting commitment to stakeholder involvement, selecting appropriate elicitation techniques and capturing requirements and scenarios in a simple to understand form. For reducing rework, teams should mature their existing requirements elicitation process by helping to define responsibilities and stakeholders, identify appropriate elicitation techniques and train team members to use the right techniques. They should also take initiative to capture user scenarios in a simple, visual form that is easy to understand by the user and accelerate elicitation with interactive simulations. Enhancement in communication and collaboration among distributed teams throughout the project lifecycle is also necessary for better requirement elicitation. Teams should also improve alignment between business expectations and project deliverables, which increases end user satisfaction and reduces rework from incorrect or incomplete requirements.

• Analysis: Reduce Time to Market with More Effective Collaboration

Requirements analysis involves verifying, estimating and prioritizing newly captured requirements for remaining application lifecycle steps. To reduce time to market, teams should codify their existing requirements analysis process by implementing an effective approach to evaluate and prioritize requirements for specification, design, construction and testing. The process is improved through the deployment of tools and

techniques that swell efficiency and accuracy. By following these steps, teams can gain better estimation and thus, superior predictability for software delivery.

 Specification: Improve Quality through More Effective Communications

Requirements specification includes adding detail to requirements incrementally to the optimal level for validation, design, coding, testing and documentation. To ameliorate quality, teams should formulate and automate their existing requirements specification process by defining a standard hierarchy of requirement types and developing standard templates to ensure completeness. They should also identify various specification techniques and apply technology for capturing requirements in a meaningful, easy-tounderstand way. Traceability across the lifecycle allows teams for better prediction of the impact of change and proactively communicates those changes to all team members affected. Through these steps, software defects can be reduced nicely because in this case development teams have a better understanding of requirements.

• Agreeing and Validation: Improve Accuracy and Completeness of Requirements

Requirements validation involves verifying whether the specification is inclusive and clear enough for the development team to understand exactly what it needed. It also includes validating whether the requirement of the key stakeholders are consistent with the original need and intent of the business. To improve accuracy, teams should mature their existing process by automating validation and verification to drive adoption and enforcement and improving consistency and quality through interactive simulations and storyboarding of visual scenarios. These steps trim down software defects, increase satisfaction and alignment with business stakeholders and thus enhance business value.

• Management: Reduce Costs through Improved Change Management

The process of gathering and managing change requests during the application lifecycle, requirements management also includes selecting changes to be incorporated within a particular release and ensuring effective implementation of changes with no adverse impact on schedule, scope or quality. To reduce costs, teams should contrive their existing requirements management process by establishing processes for defining and maintaining requirements baselines and defining a standard process for requesting changes. They may also establish a systematic approach to evaluate and approve change requests so that scope changes and

affected commitments are managed. By improving ongoing change management, maximizing business impact, while minimizing schedule and scope impact, satisfaction of business stakeholders can be increased.

VII. CONCLUSION

RE is often treated as a time-consuming, bureaucratic and contractual process and due to ineffective RE; customers may not be satisfied with the developed system. This attitude is changing as RE is increasingly recognized as a critically important activity in the field of Software Engineering. The novelty of many software applications, the speed with which they need to be developed, and the degree to which they are expected to change, all play a role in determining how the systems development process should be conducted [6]. In future, RE will continue to evolve in order to deal with different development scenarios for further amelioration. In this paper we discussed RE in detail with challenges and some suggestions. We believe that effective RE will continue to play a key role to ensure success of a project in developing quality software.

REFERENCES

- Roger S. Pressman, "Requirements Engineering," in Software Engineering: A Practitioner's Approach, McGraw-Hill, Fifth Edition, pp. 271-298, 2001.
- [2] Reifer, D.J., "Requirements Engineering," in Encyclopedia of Software Engineering (J.J. Marciniak, ed.), Wiley, 1994, pp. 1043– 1054.
- [3] Zave, P., "Classification of Research Efforts in Requirements Engineering", ACM Computing Surveys, 29(4), pp. 315-321, 1997.
- [4] Brooks, F.P. Jr. No Silver Bullet: Essence and Accidents of Software Engineering. IEEE Computer, 10-19, April 1987.
- [5] Stevens, R., Brook, P., Jackson, K. & Arnold, S. (1998). Systems Engineering: Coping with Complexity. Prentice Hall Europe.
- [6] Bashar Nuseibeh and Steve Easterbrook, Requirements Engineering: A Roadmap, In ICSE '00: Proceedings of the Conference on The Future of Software Engineering (2000), pp. 35-46.
- [7] Carter, R., Martin, J., Mayblin, B. & Munday, M. (1984). Systems, Management and Change: A Graphic Guide. London: Paul Chapman Publishing/Harper and Row.
- [8] Wieringa, R. J. (1996). Requirements Engineering: Frameworks for Understanding. Wiley.
- [9] Robertson, S. & Robertson, J. (1994). The Complete Systems Analysis: The Workbook, The Textbook, the Answers. Dorset House.
- [10] Posner, M. I. (Ed.). (1993). Foundations of Cognitive Science. MIT Press.
- [11] Goguen, J. & Jirotka, M. (Ed.). (1994). Requirements Engineering: Social and Technical Issues. London: Academic Press.
- [12] Lehman, M. M. (1980). Programs, Life Cycles, and Laws of Software Evolution. Proceedings of the IEEE, 68(9): 1060-1076.
- [13] Burg, J. F. M. (1997). Linguistic Instruments in Requirements Engineering. Amsterdam: IOS Press
- [14] Boehm, B. W. (1981). Software Engineering Economics. Englewood Cliffs, NJ: Prentice-Hall.
- [15] Nakajo, T. & Kume, H. (1991). A Case History Analysis of Software Error Cause-Effect Relationships. Transactions on Software Engineering, 17(8): 830-838.

- [16] Finkelstein, A. (1993). Requirements Engineering: an overview. 2nd Asia-Pacific Software Engineering Conference (APSEC'93), Tokyo, Japan, 1993.
- [17] Nuseibeh, B. (1997). Ariane 5: Who Dunnit? IEEE Software, 14(3): 15-16.
- [18] White Paper on "Effective Requirements Definition and Management: Improves Systems and Communication", Borland Software Corporation, May 2009.
- [19] Gottesdeiner, E., Requirements by Collaboration, Addison- Wesley, 2002
- [20] Standish Group, "The Chaos Report," www.standishgroup.com, 1995.
- [21] Hofmann, H., and F. Lehner, "Requirements Engineering as a Success Factor in Software Projects," IEEE Software, 18, 4 (July/Aug 2001), pp. 58-66.
- [22] Sharp, H., Finkelstein, A. & Galal, G. (1999). Stakeholder Identification in the Requirements Engineering Process. Workshop on Requirements Engineering Processes (REP'99) - DEXA'99, Florence, Italy, 1-3 September 1999, pp. 387-391.
- [23] Dardenne, A., Lamsweerde, A. v. & Fickas, S. (1993). Goal-Directed Requirements Acquisition. Science of Computer Programming, 20: 3-50.
- [24] Johnson, P. (1992). Human-Computer Interaction: psychology, task analysis and software engineering. McGraw-Hill.
- [25] Schneider, G. & Winters, J. (1998). Applying Use Cases: a practical guide. Addison-Wesley.
- [26] Jarke, M. & Kurki-Suonio, R. (1998). Guest Editorial Special issue on Scenario Management. IEEE Transactions on Software Engineering, 24(12).
- [27] Maiden, N. & Rugg, G. (1996). ACRE: Selecting Methods For Requirements Acquisition. Software Engineering Journal, 11(3): 183-192
- [28] Davis, A. (1992). Operational Prototyping: A New Development Approach. Software, 9(5): 70-78.
- [29] van Lamsweerde, A., Darimont, R. & Letier, E. (1998). Managing conflicts in goal-driven requirements engineering. IEEE Transactions on Software Engineering, 24(11): 908-926.
- [30] Chung, L., Nixon, B., Yu, E. & Mylopoulos, J. (2000). Non-Functional Requirements in Software Engineering. Boston: Kluwer Academic Publishers.
- [31] Maiden, N. (1998). CREWS-SAVRE: Scenarios for Acquiring and Validating Requirements. Automated Software Engineering, 5(4): 419-446
- [32] Shaw, M. & Gaines, B. (1996). Requirements Acquisition. Software Engineering Journal, 11(3): 149-165.
- [33] Goguen, J. & Linde, C. (1993). Techniques for Requirements Elicitation. 1st IEEE International Symposium on Requirements Engineering (RE'93), San Diego, USA, 4-6th January 1993, pp. 152-164.
- [34] Viller, S. & Sommerville, I. (1999). Social Analysis in the Requirements Engineering Process: from ethnography to method. 4th International Symposium on Requirements Engineering (RE'99), Limerick, Ireland, 7-11th June 1999.
- [35] Potts, C. (1997). Requirements Models in Context. 3rd International Symposium on Requirements Engineering (RE'97), Annapolis, USA, 6-10 January 1997, pp. 102-104.
- [36] Wiegers, K., Software Requirements, Microsoft Press, 1999.
- [37] Gravell, A. & Henderson, P. (1996). Executing Formal Specifications Need Not Be Harmful. IEE Software Engineering Journal, 11(2): 104-110.
- [38] Maiden, N. A. M. & Sutcliffe, A. G. (1992). Exploiting Reusable Specifications Through Analogy. Communications of the ACM, 34(5): 55-64.
- [39] Fickas, S. & Nagarajan, P. (1988). Critiquing Software Specifications: a knowledge based approach. IEEE Software, 5(6).

- [40] Holzmann, G. J. (1997). The Model Checker Spin. Transactions on Software Engineering, 23(5): 279-295.
- [41] IEEE 830-1998 Standard, found a http://standards.ieee.org/reading/ieee/std_public/description/se/830-1998_desc.html
- [42] IEEE 1233-1998 Standard, found at http://standards.ieee.org/reading/ieee/std_public/description/se/1233-1998_desc.html
- [43] Heitmeyer, C. L., Jeffords, R. D. & Labaw, B. G. (1996). Automated Consistency Checking of Requirements Specifications. IEEE Transactions on Software Engineering and Methodology, 5(3): 231-261
- [44] Jackson, M. (1995). Software Requirements and Specifications: A Lexicon of Practice, Principles and Prejudices. Addison Wesley.
- [45] Easterbrook, S. M. (1991). Resolving Conflicts Between Domain Descriptions with Computer-Supported Negotiation. Knowledge Acquisition: An International Journal, 3: 255-289.
- [46] Robinson, W. N. & Volkov, S. (1998). Supporting the Negotiation Life-Cycle. Communications of the ACM, 41(5): 95-102.
- [47] Boehm, B., Bose, P., Horowitz, E. & Lee, M. J. (1995). Requirements Negotiation and Renegotiation Aids: A Theory-W Based Spiral Approach. 17th International Conference on Software Engineering (ICSE-17), Seattle, USA, 23-30 April 1995, pp. 243-254.
- [48] Karlsson, J. & Ryan, K. (1997). Prioritizing Requirements Using a Cost-Value Approach. IEEE Software: 67-74.
- [49] Hauser, J. R. & Clausing, D. (1988). The House of Quality. The Harvard Business Review(3): 63-73.
- [50] Bohner, S. A. & Arnold, R. S. (Ed.). (1996). Software Change Impact Analysis. IEEE Computer Society Press.
- [51] Estublier, J. (2000). Software Configuration Management: A Roadmap, In ICSE '00: Proceedings of the Conference on The Future of Software Engineering (2000), pp. 279-289.
- [52] Ghezzi, C. & Nuseibeh, B. (1998). Guest Editorial Managing Inconsistency in Software Development. Transactions on Software Engineering, 24(11): 906-907.
- [53] Garlan, D. (2000). Software Architecture: A Roadmap. In ICSE '00: Proceedings of the Conference on The Future of Software Engineering (2000), pp. 91-101.

AUTHORS' PROFILE

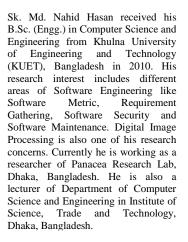
Mohammad Shabbir Hasan received his B.Sc. (Engg.) in Computer Science and Engineering from Khulna University of Engineering and Technology (KUET), Bangladesh in 2008. His research interest includes different areas of Software Engineering like Requirement Software Engineering, Metric. Software Security and Software Maintenance. He has coauthored numerous research papers published in International Journals and Conference Proceedings. Currently he is working as a researcher of Panacea Research Lab, Dhaka, Bangladesh. He is also a lecturer of Department of Computer Science and Engineering in Institute of Science, Trade and Technology, Dhaka. Bangladesh.



Abdullah Al Mahmood received his B.Sc. (Engg.) in Computer Science and Engineering from Khulna University of Engineering and Technology (KUET), Bangladesh in 2009. His research interest includes Robotics, Intelligence, Artificial Internet Security and various areas of Software Engineering like Requirement Gathering, Requirement Prioritization and Software Security. He has coauthored a good number of research papers published in International Journals and Conference Proceedings. Currently he is working as a researcher of Panacea Research Lab, Dhaka, Bangladesh. He is also a lecturer at Institute of Science, Trade and Technology, Dhaka, Bangladesh and a Project Coordinator of Technocrats



Farin Rahman received her B. Sc. (Engg.) in Computer Science and Engineering from Khulna University of Engineering and Technology (KUET), Bangladesh in 2009. Her research interest encompasses with different sectors of Software Engineering like Requirement Engineering, Software Security and Software Maintenance. She is working as a researcher in Panacea Research Lab and also as a lecturer in Daffodil Institute of Information Technology, Dhaka, Bangladesh.





GCC License Plates Detection and Recognition Using Morphological Filtering and Neural Networks

Mohamed Deriche

Dept. of Electrical Engineering King Fahd University of Petroleum & Minerals Dhahran 31261, Saudi Arabia

Abstract—License Plate Recognition (LPR) systems play an important role in intelligent transportation applications. These systems have extensively been used in highway and bridge charge, port, airport gate monitoring, parking and toll applications, to mention a few. We propose here an automatic license plate detection and recognition system for GCC countries license plates containing Arabic letters and numerals. The system introduces a robust algorithm for the extraction of the license plate region using adaptive thresholding and morphological filtering. The recognition stage is based on extracting LDA (Linear Discriminant Analysis) features with a neural network classifier. Preliminary experiments on the system have been carried with real images of vehicles captured under various conditions. The proposed system is shown to achieve high recognition accuracy under different illumination conditions.

Keywords-component; license palte recognition, LDA, Arabic character recognition; GCC countries

I. Introduction

In recent years, research on intelligent transportation systems (ITS) has gained a lot of attention. Such systems cover a multitude of technologies subdivided into intelligent infrastructure systems and intelligent vehicle systems [13]. As one of main forms of ITS technology, Automatic License Plate Recognition (ALPR) is an important technique that is used for the identification of vehicles. There are many applications that could benefit from such technology including entrance admission, security, parking control, airport or cargo control, road traffic control, speed control, toll gate automation, and so on. The two main components of any ALPR system are: plate localization & segmentation, and character segmentation & The inaccurate detection of the plate and recognition. characters leads to a useless recognition stage. On the other hand, character recognition is an essential and important step in any ALPR system, which influences significantly the overall accuracy and processing speed of the whole system [2, 14]. The problem is that most researchers either focus on plate detection or character recognition, but not on both.

localization and character segmentation, we usually need to use

some edge detection algorithm so that the final output

segmented characters are binary images. In this respect, we

propose to use morphological filtering followed by a robust

binarisation algorithm using a novel adaptive thresholding For the recognition stage, most previous work concentrated on using artificial neural networks (ANN) [16]. ANNs can achieve promising performance if the quality of the given image is good. However, the quality of images taken in real applications is not always high. This is due to the operating conditions (e.g. dust) and distortion or degradation due to poor image acquisition environment/equipment. Experiments have shown that it is difficult to achieve high recognition rates only by feeding the data from the images into the neural network [3]. In this paper, we propose to preprocess the segmented character images using a Linear Discriminant Analysis (LDA) to transform such images into small dimension feature vectors, then use these vectors as inputs to the ANN classifier. Preliminary work on the system has been carried on real images of vehicles captured under various illumination conditions with excellent overall performance compared to existing systems dealing with Arabic characters. It is worth noting that there were very few attempts, to date, in developing robust ALPR systems that support Arabic, Latin characters, and numerals.

II. BACKGROUND

Before presenting the proposed system, a brief discussion will be given on current ALPR systems for Arabic, and Latin characters, in addition to number-based car plates systems. A typical ALPR system is composed of the following steps (see Fig. 1):

A. License Plate Region Extraction

Such a step is very critical to the success of any ALPR system. Park et. al. [4] developed a model for extracting Korean license plates based on color while Kim [5] proposed a system to extract the plate based on the Hough transform method. Hontani et. al. [6] developed a method for extracting the plate without knowing its position and the image size. Their approach was based on a scale shape analysis. Ahmed et. al. [7] proposed a vertical edge detection algorithm followed by edge matching based on the size and B/W ratio of the plate. Numerous other techniques have also been proposed in the literature, see references [12] for a detailed survey of such techniques.

Fig. 1: A Typical ALPR system

B. License Plate Segmentation

Before any classification can be carried out, there is a need to segment the license plate characters from the extracted plate. Fortunately for Arabic characters based LPR systems, like in GCC countries, all the characters are in their isolated form. Exception is the country name, which can be easily isolated and dealt with separately. This makes the segmentation stage simple as can be seen in the next section [16].

C. License Plate Recognition

A number of statistical, syntactic, and neural approaches have been developed for the recognition stage. Cowell and Hussain [8] identified the characters based on the number of black pixel rows and columns of the character and comparison of the values to a set of templates or signatures in a database. Furthermore, the thinning of Arabic characters was also discussed in [9] to extract essential structural information of each character, which is then used for the classification stage. Template matching was proposed by Zidouri [7]. This approach involves the use of a database of character or templates, and for each possible input character there is a separate template. Correlation techniques are generally used to identify the difficult characters.

A structural or syntactic approach to recognize characters in a text document was adopted by Hamami [10]. This technique yields better results when applied on individual characters. However, since this approach is based on the detection of holes and concavities in the four directions (up, down, left, and right), it may not be appropriate for Arabic characters with low resolution but may work for numerals. Additionally, secondary characteristics are used in order to differentiate between the characters. Numerous systems have used neural networks for recognition [11]. Such systems as usually very effective, however their accuracy depends heavily on the size of the training data [12,17].

III. THE PROPOSED ALPR SYSTEM

The proposed system in Fig. 2 and 3 is intended for a parking application in the GCC countries. Our preliminary experiments were carried out for Saudi license plates for the moment with a plan for extending the coverage to all GCC countries and beyond.

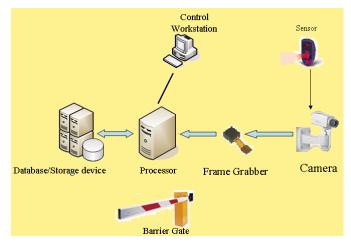


Fig. 2: The Proposed ALPR system

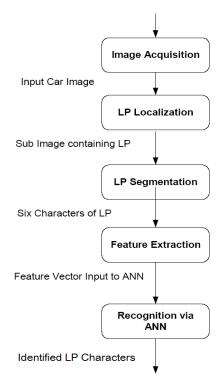


Fig. 3: Flow chart of the proposed algorithm

The system starts first by acquiring the image of the car containing the LP region (we used in our experiments frontal images). Sample images taken in our experiments are displayed in Fig. 4. Next, the license plate of the car is extracted by

applying a series of operations. The extracted LP region is further divided into individual characters. For recognition, the isolated characters are then transformed and fed into a pretrained neural network.

The sequence of recognized characters is then checked with a database. If the car is recognized, the system would then command the authorization for access or for billing or for sending the number plate to a central computer for further processing depending upon the application of interest.

The candidate LP region extraction is the key step in an LPR system, which influences the accuracy of the system significantly. The goal of this phase is, given an input image, to produce a number of regions that have high probability of containing a license plate. In the adopted approach, the extraction of the candidate region from a set of probable regions is carried out in four steps. These are explained in the following subsections.



Fig. 4: Sample images from the database

A. License Plate Extraction

The LP extraction is the main step in the overall system. The accuracy of this step affects the accuracy of the whole system significantly. This phase extracts the region of interest, i.e. the license plate from the acquired image. This step consists of three stages:

1) Edge Detection

First of all, the acquired image should be converted from a color RGB image into a gray level image (see Fig. 5). Different gradient-based edge detectors were implemented. In our experiments, the Canny edge detector was used. The Canny Edge Detector was shown to achieve low error rate, localized edge points, and a single point edge response (see Fig. 6). We

noticed from the collected data, that most car images exhibited more horizontal lines than vertical lines.





Fig. 5: The original acquired image & its gray-scale image

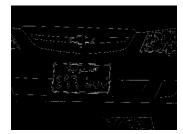




Fig. 6: Detected edges using Canny's method and edge enhancement using a thickening mask

Although the Canny's Edge Detector proved to be the best technique for detecting the LP region, an additional mask was required to enhance the detected edges. The thickening mask is used for this purpose. The thickening operation is the dual of morphological thinning and helps in strengthening the edge structure of the image (see Fig. 6).

2) Detecting the Closed Boundaries of the LP Candidate Regions

By tracing the pixels in the edge detected image, the closed structures are considered as boundaries of candidate LP regions. Tracing is carried using a 8- neighborhood scheme. After tracing the connected pixels, a set of predefined conditions is then used on these boundaries for identifying the closed contours that are LP mostly likely candidates. The conditions we use for this purpose are:

- Length and Shape Filtering: This step is based on a pixel counting approach. In our specific application, the length of the LP usually ranges between 1000 and 2400 pixels (from training). Our approach is to trace all boundaries within the given image, and select the boundaries containing between 1000 and 2400 pixels.
- Width/Length Ratio: The standard ratio of width to height in normal Saudi license plates is 2:1. To take into considerations small angle deviations, we used a ratio range of 1.3 to 2.7. As for long license plates, the ratio is around five (practically, between 3 and 5.5). If the area under study satisfies the above conditions, it is then seen as a good LP candidate.
- Black to White Ratio: Normal Saudi license plates are white with black characters. It is found that the black to

white ratio is around 20-25% while the mean is between 150 and 170 gray levels. Only regions satisfying this condition are considered as potential LP candidates. The above conditions were found to lead to a single LP region for most images from the database. (see Fig. 7).



Fig. 7: Final results from testing the different LP candidate regions

3) Filtering

Once the candidate LP region is chosen, some standard filtering tools are used to remove unwanted noise (Impulsive and Gaussian) and enhance the selected LP region.

B. License Plate Segmentation

After localizing the LP region, a robust segmentation is required in order to extract the 6 characters. Before extracting the characters, we need first to identify and remove the word "Saudia", then carry vertical and horizontal projections.

1) Removing the Word "Saudia" (السعودية)

The approach for filtering out the word "Saudia" from the LP region depends upon the type of license plate under analysis.

Case 1: Long LP: For this case, we remove the word area from the middle region on the x-axis and from the middle region on the y-axis as shown in Fig. 8. This region is then filled and replaced with the average gray level of the plate.

Case 2: Normal LP: For normal license plates, we delete the top of image containing the word "Saudia" as shown in Fig. 8.





Fig. 8: Filtering the word "Saudia" from the license plates

2) Horizontal and Vertical Projections

To obtain the separate characters, we use the horizontal and vertical projections approach. The horizontal projection is used to remove unwanted noise in the upper and lower regions that do not carry information about the characters as shown in Fig. 9. While the vertical projection aims at separating the different characters as shown in Fig. 10. A threshold is applied on the vertical and horizontal projections to obtain the individual images of the characters which are usually not of identical size.

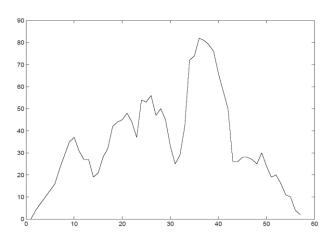


Fig. 9: The horizontal projection of the license plate

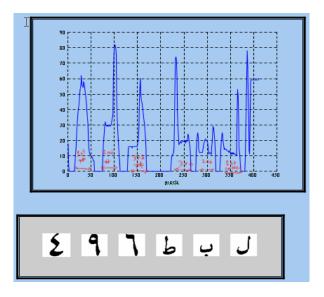


Fig. 10: The vertical projection of the license plate

C. Normalization and Character recognition

There are numerous techniques in pattern recognition that have been applied successfully to the problem of character recognition. For Arabic characters, techniques such as statistical, syntactic, and neural networks have been in use [4,12]. Most errors in these systems come from the segmentation stage. This is because Arabic writing is connected even if it is printed or typewritten. Fortunately for our application, all the characters are in their isolated form and therefore easily segmented .

Before discussing the details of the classification stage, we need to outline the preprocessing stage of the segmented characters. In particular, this first stage consists of a normalization step followed by a binarization step.

1) Normalization and Binary Images

Even though the heights of the extracted character images are all the same, the widths on the other hand may be slightly different (from the projection stage). For this reason, we normalize all the segmented imaged to 512x256 pixel images.



Fig. 11: Sample characters images after normalization and resizing

To simplify the complexity of the classification stage, we introduce two additional steps. First, all the images were resized to 1/16 of their original size resulting in images with 128x64 pixels (see Fig. 11).

Furthermore, the images were complemented to make the characters appear as white pixels over a black background. The resulting images were then converted into binary images using a novel adaptive thresholding approach. A very high threshold is initially used and the number of white pixels over the total number of pixels is determined. The threshold is then lowered in steps until a jump in this ratio is observed. The optimal threshold is then chosen as the threshold used before the jump in the value of the ratio occurs.

2) Feature Extraction and Classification using ANN

Once the images are normalized and converted into binary images, the data is not ready for processing by the classifier. Previous work with Artificial Neural Networks in classification has shown excellent results when the quality of the images used for training is high and when the training database is large enough. However, the quality of images taken in real applications is not always high. This is due to the operating

conditions (e.g. dust) and distortion or degradation due to poor image acquisition environment/equipment [12]. Previous experiments have shown that it is difficult to achieve high recognition rates only by feeding the data from the images into the neural network [3]. In this paper, we propose to preprocess the segmented character images using a Linear Discriminant Analysis (LDA) to transform such images into small dimension feature vectors, then use these vectors as inputs to the ANN classifier.

Linear discriminant analysis (LDA) is a classical dimension reduction method that aims at finding the optimal projection directions to maximize the ratio of the between-class scatter and the within-class scatter. After finding the projected directions, data can be mapped to a low-dimensional subspace. This lower dimension subspace can be seen as the subspace of feature patterns (see [18] for more details. Once the feature vectors for both the training database of characters and the test feature vectors are obtained, the classification stage using ANN is performed. Note that the size of the feature vectors is now very small (around 20x1) as compared to the size of the original images (128x64x1).

3) Classification using ANN

Artificial Neural Networks have gained a lot of popularity in different pattern recognition applications. An ANN is basically a computational model inspired from the way our biological neural system works. It uses neurons as its processing elements; working in parallel as a well as a unified network to solve learning problems. The interesting aspect of a neural network is that, like people, it learns by example and improves by training. There are different structures of ANNs. The most popular is the Multilayer Perceptron (MLP) model using the back-propagation (BP) algorithm for training. Such a network gained a lot popularity in classification tasks due to its flexibility, robustness, and computational efficiency.

The general model of the MLP consists of a number of nodes arranged in multiple layers with connections between the nodes in the adjacent layers by weights. Each of the hidden layers consists of a number of neurons. Each neuron is connected to the input via a link with a certain weight. The output of the neuron is applied to a non linear function.

If $y_i(n)$ is the output of i^{th} neuron at the n^{th} iteration, $w_{ji}(n)$ is the synaptic weight connecting the output of the i^{th} neuron to the j^{th} neuron, then the local field $v_j(n)$ induced at the input of the activation function associated with neuron j is given by:

$$v_{j}(n) = \sum_{i=0}^{m} w_{ji}(n) y_{i}(n),$$
 (1)

where m is the total number of the inputs applied to the neuron j. Let's denote the nonlinear function applied to the output of a given neuron as $\Phi(.)$, then the output at the jth neuron becomes:

$$y_j(n) = \Phi_j(v_j(n)). \tag{2}$$

The most popular nonlinear activation functions used are the "Tangent sigmoid" and the "Logarithmic sigmoid".

Using the conventional BP algorithm, the synaptic weights $w_{ii}(n)$ are updated based on the following principle:

$$w_{ji}(n+1) = w_{ji}(n) + \eta \delta_j(n) y_i(n),$$

where, η is the learning rate, and $\delta_j(n)$ is called local gradient associated with the j^{th} neuron. If the neuron j is an output neuron, then the local gradient used in the iterations is obtained from:

$$\delta_{j}(n) = e_{j}(n)\Phi_{j}(v_{j}(n)), \tag{4}$$

Where $e_j(n)$ is the error between the output of neuron $y_j(n)$ and its desired response $d_j(n)$. If the neuron j is a neuron from the hidden layer, then the local gradient according to the following expression:

$$\delta_{j}(n) = \Phi'_{j}(v_{j}(n)) \sum_{k} \delta_{k}(n) w_{kj}(n).$$
(5)

Many neural network architectures have been proposed for use in LPR systems. The most popular neural networks are the multilayer feed forward neural network, where neurons are grouped between layers and connections between neurons and consecutive layers are permitted. In this project we have used the multilayer perception.

In our application, the input characters are now represented by feature vector of dimension 20x1. Hence, the network receives the 20-element input vectors. After the training stage, the network is then used to identify the input character.

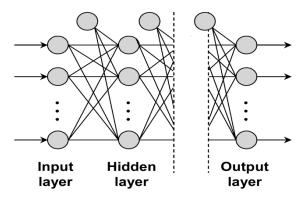


Fig. 12: Structure of an MLP network

The 27 elements of the output vector represent either a letter or a number. The Saudi LPs possible characters include 17 Arabic alphabet letters and 10 Indian numbers (see Fig. 13).



Fig. 13: Alphanumerals included in Saudi LPs (17 letters & 10 numbers)

The network is designed as a two-layer log-sigmoid/log-sigmoid network. The hidden (first) layer is chosen to have 20 neurons.

The different stages in training are:

- The network is first trained on ideal vectors until it has a predefined low sum- squared error.
- Then, the network is trained on various sets of ideal and noisy vectors. Noisy observations are used to maintain the network's ability to classify ideal input vectors.

The network is again tested on just ideal vectors. This ensures that the network responds perfectly when presented with ideal characters.

IV. EXPERIMENTAL RESULTS

We now discuss the results of the proposed ALPR system. The discussion is based on the results obtained from the previous stages. The first stage discusses the extraction technique used in the LPR system. Then, the second stage deals with the segmentation part. Next, the results of the neural network recognition technique are illustrated. Finally, the whole system performance is described.

For the sake of comparison, we also implemented the technique discussed in [13] using the AdaBoost algorithm. Both systems were trained with 500 car images and 300 other images were used for testing. We focused on the accuracy of the three stages discussed above, in particular, LP correct detection, Correct character segmentation, and Character recognition accuracy. The results are summarized in Table 1.

Table 1: THE RESULTS OF THE EXTRACTION ALGORITHM

Algorithm	Tested	Successful	Accuracy in	Character	Overall
	Plates	LP	character	Recognition	Recognition
		extraction	segmentation	Rate	Rate
Algorithm	300	278	266	263	263
from [14]		(92.7%)	(95.6%)	(98.8%)	(87.6%)
Our	300	291	287	285	285
Algorithm		(97%)	(98.6)	(99.3%)	(95%)

Vol. 8, No. 8, November 2010

The table shows clearly that the proposed system achieves excellent results at all stages of the process. The overall accuracy is defined as the product of all intermediate accuracies. The performance of the proposed system outperforms by far the results obtained in [13].

We also tested the accuracy of the system when additional noise was considered on the plates. The results have been excellent down to an SNR of 5 dB. The system is also resilient to moderate tilt in the license plate.

V. CONCLUSION

License plate recognition systems exist and commercially available for most European and Asian car However, there were very few attempts made in developing systems for Arabic characters. In this paper, we propose an ALPR system of low cost and computational complexity for Saudi Arabian license plates, and easily extendable to other GCC car plates. Beside the use of Arabic language, Saudi Arabian license plates have several unique features that are taken into account in the segmentation and recognition phases. The system was tested over a large number of car images taken under different conditions. The overall accuracy obtained was above 95%. Two major improvements were made to existing systems. First a more robust preprocessing stage was introduced which includes an adaptive thresholding approach for binarization of the images. Second, the NN classification stage was enhanced and simplified by working with LDA features extracted from the normalized character images.

VI. ACKNOWLEDGMENTS

The authors acknowledge ASTF and Abdul Latif Jameel CO. LTD for funding this work (Research Project code "IR062281"). The authors also acknowledge the support of King Fahd University of Petroleum & Minerals (KFUPM), Saudi Arabia.

REFERENCES

- [1] D. G. Bailey, D. Irecki, , B.K. Lim, and L. Yang, "Test bed for number plate recognition applications,". Proceedings of the First IEEE International Workshop on Electronic Design, Test and Applications (DELTA'02), IEEE Computer Society, 2002.
- [2] S-L. Chang, L-S. Chen, Y-C. Chung and S-W. Chen. "Automatic License Plate Recognition" IEEE Trans. on Intell. Transportation Systems, vol. 5, no. 1, pp. 42-53, 2004.
- [3] J. M. Lopez, J. Gonzalez, C. Galindo, and J. Cabello "A Versatile Lowcost Car Plate Recognition System", ISSPA07, Sharjah, February 2007.
- [4] Park, S., Kim, K., Jung, K., "Locating car license plates using neural networks", IEE electronic letters, p 1475, 1999.
- [5] Kim, G.M, "The automatic recognition of the plate of the vehicle using the correlation coefficient and Hough transform", JoC, ASE, p. 510, 1997.
- [6] Hontani, H., and T. Koga, "Character Extraction Method Without prior knowledge on size and information", Proceedings of the IEEE IVEC'01, p67, 2001.
- [7] Ahmed, M.J., Sarfraz, M., Zidouri, A., Al-Khatib, "License plate recognition system", Electronics, Circuits and Systems, 2003. ICECS 2003. Proceedings of the 2003 10th IEEE International Conference pp: 898-901 Vol.2, 14-17 Dec. 2003.

- [8] Cowell, J., and Hussain, F., "A fast recognition system for isolated Arabic characters", Proceedings of the 6th Int. conference on information and visualization, IEEE Computer Society, London, UK, 2002, p. 651.
- [9] Cowell, J., and Hussain, F., "Extracting features from Arabic characters", Proceedings of the IASTED conference on computer graphics Hawaii, 2001, p. 201.
- [10] Hamami, L. and Berkani, D., "Recognition System for Printed Multifont and Multi-size Arabic Characters", AJSE, 2002.
- [11] Chang, S., Li-Shien Chen, Yun-Chung Chung and Sei-Wan Chen, "Automatic license plate recognition", Intelligent Transportation Systems, IEEE Transactions on Volume 5, Issue 1, p:42 – 53, March 2004.
- [12] Anagnostopoulos, C.-N.E.; Anagnostopoulos, I.E.; Psoroulas, I.D.; Loumos, V.; Kayafas, E.; , "License Plate Recognition From Still Images and Video Sequences: A Survey," Intelligent Transportation Systems, IEEE Transactions on , vol.9, no.3, pp.377-391, Sept. 2008.
- [13] L. Zheng, X Hu, B. Samali, L. Yang, "Accuray enhancement of license plate recognition", Proc. Of Int Conf Computer and Information Technology, CIT, 2010
- Technology, CIT, 2010
 [14] B. Shen," License plate character segmentation and recognition based on RBFNN", 2nd Workshop on Education Technology and Computer Science, 2010.
- [15] H. Mahin, S. Kasaei, F. Dorri, "An efficient feature based license plate localization method", Proc. ICPR, 2006.
- [16] A. Zidouri, M. Deriche, Recognition of Arabic License Plates using NN, First Workshop on Image Processing Theory, Tools and Applications, IPTA 2008.
- [17] S. Qiao, Y. Zha, X. Li, T. Liu, B. Zhang, "Research on improving the accuracy of license plate character segmentation", 5th Int Conf on Frontiers of Computer Science and Technology, 2010
- [18] K. Etemad, R. Chellappa, Discriminant Analysis for Recognition of Human Face Images, Journal of the Optical Society of America A, Vol. 14, No. 8, August 1997, pp. 1724-1733.

Combined Algorithm of Particle Swarm Optimization

Narinder Singh Department of Mathematics, Punjabi University, Patiala INDIA,(Punjab)-147201

S.B. Singh Department of Mathematics, Punjabi University, Patiala, INDIA,(Punjab)-147201

J.C.Bansal ABV-Indian Institute of Information Technology and Management-Gwalior (M.P), INDIA

Abstract: A new optimization algorithm is developed in this paper as a Combined Algorithm of particle swarm optimization, is presented, based on a novel philosophy by modifying the velocity update equation. This is done by combined two different PSO algorithms i.e., Standard Particle Swarm Optimization and Personal Best Position **Particle** Swarm Optimization. performance is compared with the standard PSO (SPSO) by testing it on a set 15 of scalable and 13 non-scalable test problems. Based on the numerical and graphical analyses of results it is shown that the CAPSO (Combined Algorithm of Particle Swarm Optimization) outperforms the (Standard Particle Swarm Optimization), in terms of efficiency, reliability, accuracy and stability.

Keywords: Particle Swarm Optimization, CAPSO (Combined Algorithm of Particle Swarm Optimization), global optimization, velocity update equation, Personal Best Position Particle Swarm Optimization.

I. INTRODUCTION

Standard Particle Swarm Optimization (SPSO): Particle swarm optimization (PSO) [1], [2] is a stochastic, population-based search method, modeled after the behavior of bird flocks. A PSO algorithm maintains a swarm of individuals (called particles), where each individual (particle) represents a candidate solution. Particles follow a very simple behavior: emulate the success of neighboring particles, and own successes achieved. The position of a particle is therefore influenced by the best particle in a neighborhood, as well as the best solution found by the particle. Particle position x_i are adjusted using

$$x_i(t+1) = x_i(t) + v_i(t+1)$$
 ...(1)

where the velocity component, $v_i(t)$ represents the step size. For the basic PSO.

$$v_{ij}(t+1) = v_{ij}(t) + c_1 r_{ij}(y_{ij} - x_{ij}) + c_2 r_{2i}(\hat{y}_i - x_{ij}) \qquad \dots \qquad (2)$$

where w is the inertia weight [11], c_1 and c_2 are the acceleration coefficients, $r_{ij}, r_{2j} \sim U(0,1)$, \mathcal{Y}_{ij} is the personal best position of particle i, and $\hat{\mathcal{Y}}_j$ is the neighborhood best position of particle i. The neighborhood best position \mathcal{Y}_i , of particle i depends on the neighborhood topology used [3], [4]. If a star topology is used, then $\hat{\mathcal{Y}}_i$ refers to the best position found by the entire swarm. That is,

 $y_i \sim \{y_0(t), y_1(t), ..., y_s(t)\} = \min(f(y_0(t)), f(y_1(t)), ..., f(y_s(t))$ where s is the swarm size. The resulting algorithm is referred to as the global best PSO. For the ring topology, the swarm is divided into overlapping neighborhoods of particles. In this case, \hat{y}_i is the best position found by the neighborhood of particle i. The resulting algorithm is referred to as the Local best PSO.

The Von Neumann topology defines neighborhoods by organizing particles in a lattice structure. A number of empirical studies have shown that the Von Neumann topology outperforms other neighborhood topologies [4], [5]. It is important to note that neighborhoods are determined using particle indices, and are not based on any spatial information.

A large number of PSO variations have been developed, mainly to improve the accuracy of solutions, diversity, and convergence behavior [6], [7]. This section reviews those variations used in this study, from which concepts have been borrowed to develop a new, parameter-free PSO algorithm.

Van den Bergh and Engelbrecht [8], [9], and Clerc and Kennedy [3], formally proved that each particle converges to a weighted average of its personal best and neighborhood best positions. That is,

$$\lim_{x \to \infty} x_{ij} = \frac{c_1 y_{ij} + c_2 \hat{y}_{ij}}{c_1 + c_2} \qquad \dots (3)$$

This theoretically derived behavior provides support for the barebones PSO developed by Kennedy [10], where the velocity vector is replaced with a vector of random numbers sampled from a Gaussian distribution with the mean defined by equation (3), assuming that $c_1 = c_2$, and deviation,

$$\sigma = \left| y_{ij} - \hat{y}_{ij} \right|$$

The velocity equation changes to

$$v_{ij}(t+1) \sim N\left(\frac{y_{ij}+\hat{y}_{ij}}{2},\sigma\right)$$

The position update then changes to

$$x_i(t+1) = v_i(t+1)$$

Kennedy [9] also proposed an alternative version of the barebones PSO, where

$$v_{ij}(t+1) = \begin{cases} y_{ij} & if \ U(0,1) < 0.5 \\ N(\frac{y_{ij} + \hat{y}_{ij}}{2}, \sigma) & otherwise \end{cases} ..(4)$$

Based on equation (4), there is a 50% chance that the j-th dimension of the particle dimension changes to the corresponding personal best position. This version of the barebones PSO biases towards exploiting personal best positions.

Silva et al. (2002) presented a predatorpray model to maintain population diversity.

II. THE PROPOSED COMBINED ALGORITHM OF PSO

The motivation behind introducing CAPSO is that in the velocity update equation instead of comparing the combined two difference PSO algorithm update velocity equation i.e., SPSO (Standard Particle Swarm Optimization) and PBPPSO (Personal Best Position Particle Swarm Optimization).

Thus, we introduce a new velocity update equation as fellows:

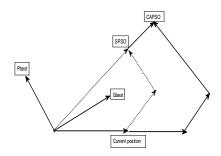
$$\begin{split} v_{ij}(t+1) &= w v_{ij}(t) + c_1 r_{1j}(y_{ij} - x_{ij}) + c_2 r_{2j}(\hat{y}_j - x_{ij}) + \\ w v_{ij}(t) &+ c_1 r_{1j}(y_{ij} - x_{ij}) + c_2 r_{2j}(-x_{ij}) \end{split}$$

OR

$$v_{ij}(t+1) = 2wv_{ij}(t) + 2c_1r_{1j}(y_{ij} - x_{ij}) + c_2r_{2j}(\hat{y}_j - 2x_{ij})$$
 (5)

In the velocity update equation of this new PSO the first term represents the current velocity of the particle and can be thought of as a momentum term. The second term is proportional to the vector $2c_1r_{1j}(y_{ij}-x_{ij})$, is responsible for the attractor of particle's current position and positive direction of its own best position (pbest). The third term is proportional to the vector $c_2r_{2j}(\hat{y}_j-2x_{ij})$, is responsible for the attractor of particle's current position.

Figure:-I: Comparative movement of a particle in SPSO and CPSO



The pseudo code of CAPSO is shown below:

ALGORITHM- CAPSO

For t = 1 to the max: bound of the number on iterations,

For i=1 to the swarm size,

For i=1 to the problem dimensionality,

Apply the velocity update equation (5);

Update Position using equation (2);

End-for- i

Compute fitness of updated position; If needed, update historical information for personal best position and global best position; End-for-i;

Terminate if global best position meets problems requirements;

End-for-t;

END ALGORITHM

III. THE TEST BED

Many times it is found that the evaluation of a proposed algorithm is evaluated only on a few benchmark problems. However, in this paper we consider a test bed of thirty benchmark problems with varying difficulty levels and problem size. The relative performance of SPSO and CAPSO is evaluated on two kinds of problem sets. Problem Set 1 consists of 15 scalable problems, i.e., those problems in which the dimension of the problems can be increased / decreased at will.

In general, the complexity of the problem increases as the problem size is increased. Problem Set 2 consists of those problems in which the problem size is fixed, but the problems have many local as well as global optima. The Problem Set 1 is shown in Table 1 and Problem Set 2 is shown in Table 2.

Table-1: Details of Problem Set-I (Continued)

Serial No	Function Name	Expression	Search Space	Objective Function Value
1.	Ackley	$Min f(x) = -20 \exp(-0.02 \sqrt{n^{-1} \sum_{i=1}^{n} x_i^2})$	$-30 \le x_i \le 30$	0
		$-\exp(n^{-1} \sum_{i=1}^{n} \cos(\pi x_{i})) + 20 + e$		
2.	Cosine Mixture	$Min f(x) = -0.1 \sum_{i=1}^{n} \cos(5\pi x_i) + \sum_{i=1}^{n} x_i^2$	$-1 \le x_i \le 1$	-0.1×(n)
3.	Exponential	$Min f(x) = (-0.5 \sum_{i=1}^{n} x_i^2)$	$-1 \le x_i \le 1$	-1
4.	Griewank	$Min f(x) = 1 + \frac{1}{4000} \sum_{i=1}^{n} x_i^2 - \prod_{i=1}^{n} \cos(\frac{x_i}{\sqrt{i}})$	$-600 \le x_i \le 600$	0
5.	Rastrigin	$Min f(x) = 10n + \sum_{i=1}^{n} [x_i^2 - 10\cos(2\pi x_i)]$	$-5.12 \le x_i \le 5.12$	0
6.	Function '6'	$Min f(x) = \sum_{i=1}^{n-1} \left[100(x_{i+1} - x_i^2)^2 + (x_i - 1)^2 \right]$	$-30 \le x_i \le 30$	0
7.	Zakharov's	$Min f(x) = \sum_{i=1}^{n} x_i^2 + \left[\sum_{i=1}^{n} (\frac{i}{2})x_i\right]^2 + \left[\sum_{i=1}^{n} (\frac{i}{2})x_i\right]^4$	$-5.12 \le x_i \le 5.12$	0
8.	Sphere	$Min f(x) = \sum_{i=1}^{n} x_i^2$ $M in f(x) = \sum_{i=1}^{n} i x_i^2$	$-5.12 \le x_i \le 5.12$	0
9.	Axis parallel hyper ellipsoid	$M \ in \ f(x) = \sum_{i=1}^{n} ix_{i}^{2}$	$-5.12 \le x_i \le 5.12$	0
10.	Schwefel '3'	$M in f(x) = \sum_{i=1}^{n} x_{i} + \prod_{i=1}^{n} x_{i} $	$-10 \le x_i \le 10$	0
11.	Dejong	$M in f(x) = \sum_{i=1}^{n} (x_i^4 + rand(0,1))$	$-10 \le x_i \le 10$	0
12.	Schwefel '4'	$Min f(x) = Max\{ \left x_i \right , 1 \le i \le n \}$	$-100 \le x_i \le 100$	0
13.	Cigar	$M \text{ in } f(x) = x_i^2 + 100000 \sum_{i=1}^n x_i^2$	$-10 \le x_i \le 10$	0
14.	Brown '3'	$Min \ f(x) = \sum_{i=1}^{n-1} \left[(x_i^2)(x_{i+1}^2 + 1) + (x_{i+1}^2 + 1)(x_i^2 + 1) \right]$	$-1 \leq x_i \leq 4$	0
15.	Function '15'	$M \text{ in } f(x) = \sum_{i=1}^{n-1} \left[0.2 x_i^2 + 0.1 x_i^2 \sin 2 x_i \right]$	$-10 \le x_i \le 10$	0

Table 2: Details of Problem Set-II

Serial No	Function Name	Expression	Search Space	Objective Function Value
1.	Becker and Lago	$M in f(x) = (x_1 - 5)^2 + (x_2 - 5)^2$	$-10 \le x_i \le 10$	0
2.	Bohachevsky '1'	$Min \ f(x) = x_1^2 + 2x_2^2 - 0.3\cos(3\pi x_1) - 0.4\cos(4\pi x_2) + 0.7$	$-50 \le x_1, x_2 \le 50$	0

3.	Bohachevsky '2'	$Min f(x) = x_1^2 + 2x_2^2 - 0.3\cos(3\pi x_1)\cos(4\pi x_2) + 0.3$	$-50 \le x_1, x_2 \le 50$	0
4.	Branin	$Min f(x) = a(x_2 - bx_1^2 + cx_1 - d)^2 + g(1 - h)\cos(x_1) + g$ $a = 1, b = \frac{5 \cdot 1}{4\pi^2}, c = \frac{5}{\pi}, d = 6, g = 10, h = \frac{1}{8\pi}$	$-5 \le x_1 \le 100$ $-5 \le x_2 \le 15$	0.398
5.	Eggcrate	$Min f(x) = x_1^2 + x_2^2 + 25(\sin^2 x_1 + \sin^2 x_2)$	$-2\pi \leq x_i \leq 2\pi$	0
6.	Miele and Cantrell	$Min f(x) = (\exp(x_1) - x_4)^4 + 100(x_2 - x_3)^6 + (\tan(x_3 - x_4))^4 + x_1^8$	$-1 \le x_i \le 1$	0
7.	Modified Rosenbrock	$Min f(x) = 100(x_2 - x_1^2)^2 + [6.4(x_2 - 0.5)^2 - x_1 - 0.6]^2$	$-5 \le x_1, x_2 \le 5$	0
8.	Easom	$Min f(x) = -\cos(x_1)\cos(x_2)\exp(-(x_1 - \pi)^2 - (x_2 - \pi)^2)$	$-10 \le x_i \le 10$	-1
9.	Periodic	$Min f(x) = -1 + \sin^2 x_1 + \sin^2 x_2 - 0.1 \exp(-x_1^2 - x_2^2)$	$-10 \le x_i \le 10$	0.9
10.	Powell's	$Mnf(x) = (x_1 + 10x_2)^2 + 5(x_3 - x_4)^2 + (x_2 - 2x_3)^4 + 10(x_1 - x_4)^4$	$-10 \le x_i \le 10$	0
11.	Camel back-3	$Min \ f(x) = 2x_1^2 + 1.05x_1^4 + \frac{1}{6}x_1^6 + x_1x_2 + x_2^2$	$-5 \le x_1, x_2 \le 5$	0
12.	Camel back-6	$Min f(x) = 4x_1^2 + 2.1x_1^4 + \frac{1}{3}x_1^6 + x_1x_2 - 4x_2^2 + 4x_2^4$	$-5 \le x_1, x_2 \le 5$	-1.0316
13.	Aluffi-Pentini's	$Min f(x) = 0.25x_1^4 - 0.5x_1^4 - 0.5x_1^2 + 0.1x_1 + 0.5x_2^2$	$-10 \le x_i^- \le 10$	-0.3523

IV. ANALYSES OF RESULTS

The SPSO and the CAPSO are coded in C++ and implemented on Pentium-IV 2.4 GHz machine with 512 MB RAM under WINXP platform. Thirty independent runs with different seed for the generation of random numbers are taken. However, the same seed is used for generating the initial swarm for SPSO and CAPSO for the i^{th} run, where i = 1, 2, ..., 50.

A run is said to be a successful run if the best objective function value found in that run lies within 1% accuracy of the best known objective function value of the problem.

The maximum number of function evaluations is fixed to be 30,000. The swarm size is fixed to 20 and dim is 30. The inertia weight is 0.7 and the acceleration coefficients for SPSO and CAPSO are set to be $c_1 = c_2 = 1.5$.

A number of criterions are used to evaluate the performance of SPSO and CAPSO.

The percentage of success is used to evaluate the reliability. The average number of function evaluations of successful runs and the average

computational time of the successful runs, are used to evaluate the cost. For problem SET-I, by fixing for problem measured by the minimum, mean, success of rate and standard deviation of the objective function values out of fifty runs. This is shown in Table 3. The corresponding information for problem SET-II is shown Table 4, respectively.

In observing Table 3, it can be seen that CAPSO gives a better quality of solutions as compared to SPSO. Thus, for the scalable problems CAPSO outperforms SPSO with respect to efficiency, reliability, cost and robustness.

In observing Table 4, it can be seen that CAPSO gives a better quality of solutions as compared to SPSO. Thus, for the non-scalable problems CAPSO outperforms SPSO with respect to efficiency, reliability, cost and robustness.

In Table 3, It is observed that SPSO could not solve two problems with 100% success, whereas CAPSO solved all the problems with 100% success.

Table-3 Comparative Objective function value obtained in 50 runs by SPSO and CAPSO for problem Set-I

Problem	Minimum	Function	Mean Funct	ion Value	Standard D	eviation	Rate of Success	
No.	Va	lue						
	SPSO	CAPSO	SPSO	CAPSO	SPSO	CAPSO	SPSO	CAPSO
1	0.667619	0.271435	16485.6000	2331.0000	0.142795	0.139467	98.00%	100%
2	0.644392	0.279915	1708.20000	193.20000	0.053545	0.129501	100%	100%
3	0.000000	0.000000	60.000000	60.000000	0.000207	0.000080	100%	100%
4	0.777974	0.422415	14364.6000	5998.8000	0.026005	0.117862	100%	100%
5	27.127816	0.266949	30000.0000	19504.800	29.809592	41.94101	0.00%	100%
6	0.000061	0.000001	166.200000	141.00000	0.200616	0.285178	100%	100%
7	0.000274	0.000003	72.000000	76.800000	0.229660	0.233736	100%	100%
8	0.685057	0.295695	6096.00000	864.00000	0.054336	0.155304	100%	100%
9	0.000002	0.000001	60.600000	63.600000	0.179978	0.219324	100%	100%
10	0.001109	0.001107	60.600000	60.600000	0.161759	0.180857	100%	100%
11	0.601870	0.098892	11341.8000	5126.4000	0.067786	0.227513	100%	100%
12	0.022248	0.006012	78.000000	87.000000	0.243564	0.253503	100%	100%
13	0.001848	0.001248	1767.00000	1767.0000	0.253535	0.273535	100%	100%
14	0.000126	0.000108	60.000000	60.000000	0.048579	0.053570	100%	100%
15	0.000009	0.000001	60.000000	60.000000	0.005729	0.004014	100%	100%

Table-4 Comparative Objective function value obtained in 50 runs by SPSO and CAPSO for problem Set-II

Problem	Minimum	Function	Mean Function	on Value	Standard I	Deviation	Success of Rate	
No.	Value							
	SPSO	CAPSO	SPSO	CAPSO	SPSO	CAPSO	SPSO	CAPSO
1	0.500000	0.500000	60.000000	60.000000	0.042453	0.042452	100%	100%
2	0.017193	0.002786	64.200000	62.400000	0.258362	0.248258	100%	100%
3	0.001029	0.001024	66.600000	67.800000	0.224219	0.257928	100%	100%
4	0.398600	0.395682	128.400000	175.200000	0.137710	0.148115	100%	100%
5	0.018613	0.002431	72.000000	72.600000	0.240972	0.221812	100%	100%
6	0.498600	0.398600	128.400000	120.400000	0.167710	0.147710	100%	100%
7	0.027193	0.017786	64.200000	62.400000	0.358362	0.288258	100%	100%
8	0.015341	0.012461	82.200000	95.400000	0.281294	0.256433	100%	100%
9	0.480507	0.480489	60.000000	60.000000	0.026709	0.021144	100%	100%
10	0.067997	0.051277	840.600000	517.200000	0.215576	0.253873	100%	100%
11	0.003378	0.002978	60.600000	64.600000	0.207517	0.246517	100%	100%
12	0.005549	0.003824	63.600000	66.600000	0.270722	0.238520	100%	100%
13	0.002655	0.002017	65.400000	60.000000	0.229666	0.181436	100%	100%

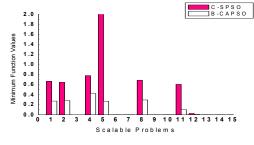


Figure A: Comparing the SPSO and CAPSO with the help of Scalable 15 Problems SET-I.

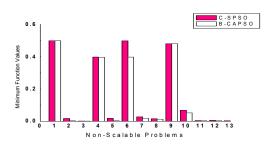


Figure B: Comparing the SPSO and CAPSO with the help of Non-Scalable 13 Problems SET-II.

V. CONCLUSIONS

This paper presented new population based algorithm CPSO (Combined Algorithm of Particle Swarm Optimization). It is represent based on a combined two different particle swarm optimization algorithms i.e., Standard Particle Swarm Optimization and Personal Best Position Particle Swarm Optimization. It is tested on 15 scalable problems and 13 nonscalable problems. It is shown that the new CAPSO (Combined Algorithm of Particle Optimization) outperforms SPSO (Standard Particle Swarm Optimization) in terms of efficiency, accuracy, reliability and robustness. Particularly for large size problems CAPSO outperforms SPSO. In this paper the effect of change of parameters in CAPSO is not explored. In a future study parameters fine tuning may be carried out for better performance. Also the application of CAPSO to the real word problems would be interesting as a future research.

REFERENCES

- [1] R.C. Eberhart and J. Kennedy. A New Optimizer using Particle Swarm Theory. In Proceedings of the Sixth International Symposium on Micromachine and Human Science, pages 39–43, 1995.
- [2] J. Kennedy and R.C. Eberhart. Particle Swarm Optimization. In Proceedings of the IEEE International Joint Conference on Neural Networks, pages 1942–1948. IEEE Press, 1995.
- [3] J. Kennedy. Small Worlds and Mega-Minds: Effects of Neighborhood Topology on Particle Swarm Performance. In Proceedings of the IEEE Congress on Evolutionary Computation, volume 3, pages 1931–1938, July 1999.
- [4] J. Kennedy and R. Mendes. Population Structure and Particle Performance. In Proceedings of the IEEE Congress on Evolutionary Computation, pages 1671– 1676. IEEE Press, 2002.
- [5] E.S. Peer, F. van den Bergh, and A.P. Engelbrecht. Using Neighborhoods with the Guaranteed Convergence PSO. In Proceedings of the IEEE Swarm Intelligence Symposium, pages 235–242. IEEE Press, 2003.
- [6] A.P. Engelbrecht. Fundamentals of Computational Swarm Intelligence. Wiley & Sons, 2005.
- [7] J. Kennedy, R.C. Eberhart, and Y. Shi. Swarm Intelligence. Morgan Kaufmann, 2001.

- [8] F. van den Bergh. An Analysis of Particle Swarm Optimizers. PhD thesis, Department of Computer Science, University of Pretoria, Pretoria, South Africa, 2002.
- [9] F. van den Bergh and A.P. Engelbrecht. A Study of Particle Swarm Optimization Particle Trajectories. Information Sciences, 176(8):937–971, 2006.
- [10] J. Kennedy. Bare Bones Particle Swarms. In Proceedings of the IEEE Swarm Intelligence Symposium, pages 80–87, April 2003.
- [11] Y. Shi and R.C. Eberhart. A Modified Particle Swarm Optimizer. In Proceedings of the IEEE Congress on Evolutionary Computation, pages 69–73, May 1998.
- [12] Angline, P.J.(1998a) 'Evolutionary optimization versus particle swarm optimization philosophy and performance differences', Lecture Notes in Computer Science, Vol.1447,pp.601-610, Springer, Berlin.
- [13] Angline, P.J(1998b) 'Using selection to improve particle swarm optimization', Proceedings of the IEEE Conference on Evolutionary Computations, pp.84-89.
- [14] Banks, A., Vincent, J. and Anyakoha, C. (2007) 'A review of particle swarm optimization,Part I:Background and development', Natural Computing: an International Journal, Vol. 6, No. 4, pp.467–484.
- [15] Banks, A., Vincent, J. and Anyakoha, C. (2008) 'A review of particle swarm optimization, Part II: Hybridisation, combinatorial, multicriteria and constrained optimization and indicative applications', Natural Computing: an International Journal, Vol. 7, No. 1, pp.109–124.
- [16] Baskar, S. and Suganthan, P.M. (2004) 'A novel concurrent particle swarm optimization', Proceedings of the Congress on Evolutionary Computations, pp.792–796.
- [17] Deep, K. and Thakur, M. (2007) 'A new crossover operator for real coded genetic algorithms', Applied Mathematics and Computation, Vol. 188, No. 1, pp.895–911. Eberhart, R.C. and Shi, Y. (2000) 'Comparing inertia weights and constriction factors in particle swarm optimization', Proc. Congress on Evolutionary Computation, San Diego, CA, pp.84–88,.
- [18] Esquivel, S.C. and Coello Coello, C.A. (2003) 'On the use of particle swarm

- optimization with multi modal functions', Proceedings of the Congress on Evolutionary Computations, pp.1130–1136.
- [19] He, S., Wu, Q.H., Wen, J.Y., Saunders, J.R. and Paton, R.C. (2004) 'A particle swarm optimizer with passive congregation', Biosystems, Vol. 78, pp.135–147.
- [20] Hendtlass, T. (2003) 'Preserving diversity in particle swarm optimization', Lecture Notes in Computer Science, Vol. 2718, pp.31–40. Higasbi, N. and Iba, H. (2003) 'Particle swarm optimization with Gaussian mutation', Proceedings of the 2003 IEEE Swarm Intelligence Symposium, pp.72–79.
- [21] Hu, X., Shi, Y. and Eberhart, R.C. (2004) 'Recent advances in particle swarm', Proceedings of Congress Evolutionary Computation, Vol. 1, pp.90–97.
- [22] Janson, S. and Middendorf, M. (2005) 'A hierarchical particle swarm optimizer and its adaptive variant', IEEE Transaction on System, Man and Cybernetics, Part B, Vol. 38, pp.1272–1282.
- [23] Krink, T. and Lovbjerg, M. (2002) 'The lifecycle model: combining particle swarm optimization, genetic algorithms and hill climbing', Proceedings of Parallel Problem solving from Nature, Vol. 7, pp.621–630.
- [24] Krink, T., Vesterstrem, J.S. and Riget, J. (2002) 'Particle swarm optimization with spatial particle extension', Proceedings of the Congress on Evolutionary Computation, pp.1474–1479.
- [25] Liang, J.J., Qin, A.K., Suganthan, P.N. and Baskar, S. (2004) 'Particle swarm optimization algorithms with novel learning strategies', Proceedings of the IEEE Conference on Systems, Man and Cybernetics, pp.3659–3664.
- [26] Liu, H., Li, B., Wang, X., Ji, Y. and Tang, Y. (2004) 'Survival density particle swarm optimization for neural network training', ISNN (1), LNCS 3173, pp.332–337, Springer-Verlag.
- [27] Lovbjerg, M. and Krink, T. (2002) 'Extending particle swarm optimizers with self-organized criticality', Proceedings of the Congress on Evolutionary Computation, pp. 1588–1593.
- [28] Lovbjerg, M., Rasmussen, T.K. and Krink, T. (2001) 'Hybrid particle swarm optimizer with breeding and subpopulation', Proceedings of the Third Genetic and Evolutionary Computation Conference, pp.469–476.

- [29] Mohais A., Ward, C. and Posthoff, C. (2004) 'Randomized directed neighborhood with edge migration in particle swarm optimization', Proceedings of the IEEE Conference on Evolutionary Computational, pp.548–555.
- [30] Pasupuleti, S. and Battiti, R. (2006) 'The gregarious particle swarm optimizer (G-PSO)',
 Proceedings of the 8th Annual Conference on Genetic and Evolutionary Computation (Seattle, Washington, USA). GECCO '06,
- [31] Peram, T., Veeramachaneni, K. and Mohan, C. K. (2003) 'Fitness-distance-ratio based particle swarm optimization', Proceedings of the IEEE Swarm Intelligence Symposium, pp.174–181.

pp.67-74, ACM, New York, NY.

- [32] Poli, R., Laugdon, W.B. and Holland, O. (2005) 'Extending particle swarm optimization via genetic programming', Proceedings of the Eighth European Conference on Genetic Programming, pp.291–300.
- [33] Riget, J. and Vesterstrem, J.S. (2002) 'A diversity-guided particle swarm optimizer the ARPSO', Technical Report 2002-02, EVALife, Department of Computer Science, University of Aarbus.
- [34] Shi, Y. and Eberhart, R.C. (1998a) 'A modified particle swarm optimizer', Proceedings of the IEEE International Conference on Evolutionary Computation, pp.69–73.
- [35] Shi, Y. and Eberhart, R.C. (1998b)
 'Parameter selection in particle swarm optimization,' 7th Annual Conference on Evolutionary Programming, San Diego, USA.
- [36] Y.Fukuyama, "Parctical Equipment Models for Fast Distribution Power Flow Considering Interconnection of Distributed Generators", Pro. Of IEEE PES Summer Meeting, 2001.
- [37] F.F. Wu and A.F. Neyer, "Asynchronous Distributed State Estimation for Power Distribution Systems", Proc. of 10th PSCC, Aug. 1990.

Optimal Solution of 2-Dim Rectangle Packing Problem based on Particle Swarm Optimization

Narinder Singh Department of Mathematics, Punjabi University, Patiala INDIA-47201

S.B. Singh Department of Mathematics, Punjabi University, Patiala, INDIA,-147201

J.C.Bansal ABV-Indian Institute of Information Technology and Management-Gwalior (M.P), INDIA

Abstract-In the rectangle-packing problem, rectangular part are placed rectangular stock sheet. Which is bigger in size in comparison to items with the aim of minimizing the unused space. This problem belongs to class of NP-complete problems where computation time for an exact solution increases with N (total items considered in the problem) and become rapidly prohibitive in cost as N increases. The solution approach to these problems lies in finding in optimal solutions while reducing the exhaustive search of all possible arrangements of nesting the parts and subsequently checking upon the execution time. Usually, various heuristic rule are proposed to generate different patterns that are near optimal. These heuristics are generally the priority rules used to allocate patterns to the stock sheet sequentially.

In this paper, the optimal solution of rectangle-packing find by Particle Swarm Optimization. Nee (1991) has been used to generate 120 feasible patterns with different sheet utilization factors. The algorithm can be

I. INTRODUCTION

Background: There are so many small scale industries/business houses which are operating in our local housing colonies. They are the trader building doors, windows and other furniture items; tailors putting up all sewing patterns on cloth material; supplier who provide marble slabs for flooring in houses, where dimensions are rarely regular, or, a cobbler determined for cutting maximum number of shoes or gloves from a piece of leather. On the other hand, we have medium scale hosiery firms of Ludhiana (India), which are producing shirts, jeans and other articles, or, capital intensive industries, like transformer division of M/s Crampon Greaves Limited at Mumbai and Bhopal, where 50 tons

used for different object sizes available in varying quantity each having many feasible patterns so as to meet the demand for items in a holistic manner. Also, the solution can be obtained by solving the LPP that takes in most viable solution generated by the revised heuristic, requirements of shapes and availability of sheets.

In the present study a comparison has been made that makes use of LPP to look for solution Vs. gradually meeting the demand by taking optimal feasible solutions generated by the heuristic that gives the best sheet utilization and exhausting the demand at each step till the demand is fully met.

Keywords: Cutting and Packing, Rectangle packing, Nesting, Assortment Problems, Heuristics, NP-complete problems, Scrap Management, Sheet Layout, Optimization Techniques, Particle Swarm Optimization.

of sheets annually are used towards the fabrication of transformers. What's more, the resource sheet can be a motherboard that is few centimeters in dimensions and looks for so many chips to be laid on it or it can be a huge rectangular sheet from which brackets/floors are to be cut in pairs that are generally symmetric about a line for submarines. The share aims of all these stated examples is to conserve resource material as scrap left-over is not recycled or reused, but treated as wastage. Arranging shapes on resource material is the critical issue that results in ultimate utilization of the resource material while observing the manufacturing constraints specific to an industrial situation. So, they all strive hard by adopting various

techniques to conserve the depleting resources available to them in their respective fields and operate in the market for profit. This problem is classified as the problem of Nesting by the researchers.

Particle Swarm Optimization: Particle swarm optimization (PSO) [1], [2] is a stochastic, population-based search method, modeled after the behavior of bird flocks. A PSO algorithm maintains a swarm of individuals (called particles), where each individual (particle) represents a candidate solution. Particles follow a very simple behavior: emulate the success of neighboring particles, and own successes achieved. The position of a particle is therefore influenced by the best particle in a neighborhood, as well as the best solution found by the particle. Particle position x_i are adjusted using

$$x_i(t+1) = x_i(t) + v_i(t+1)$$
 ..(1)

where the velocity component, $v_i(t)$ represents the step size. For the basic PSO.

 $v_{ij}(t+1) = v_{ij}(t) + c_1 r_{1j}(y_{ij} - x_{ij}) + c_2 r_{2j}(\hat{y}_j - x_{ij})$..(2) where w is the inertia weight [11], c_1 and c_2 are the acceleration coefficients, $r_{1j}, r_{2j} \sim U(0,1)$, y_{ij} is the personal best position of particle i, and \hat{y}_j is the neighborhood best position of particle i.

The neighborhood best position y_i , of particle i depends on the neighborhood topology used [3], [4]. If a star topology is used, then \hat{y}_i refers to the best position found by the entire swarm. That is,

 $y_i \sim \{y_0(t), y_1(t), ..., y_s(t)\} = \min(f(y_0(t)), f(y_1(t)), ..., f(y_s(t))$ where s is the swarm size. The resulting algorithm is referred to as the global best (gbest) PSO. For the ring topology, the swarm is divided into overlapping neighborhoods of particles. In this case, \hat{y}_i is the best position found by the neighborhood of particle i. The resulting algorithm is referred to as the Local best (lbest) PSO.

The Von Neumann topology defines neighborhoods by organizing particles in a lattice structure. A number of empirical studies have shown that the Von Neumann topology outperforms other neighborhood topologies [4], [5]. It is important to note that neighborhoods are

determined using particle indices, and are not based on any spatial information.

A large number of PSO variations have been developed, mainly to improve the accuracy of solutions, diversity, and convergence behavior [6], [7]. This section reviews those variations used in this study, from which concepts have been borrowed to develop a new, parameter-free PSO algorithm.

Van den Bergh and Engelbrecht [8], [9], and Clerc and Kennedy [3], formally proved that each particle converges to a weighted average of its personal best and neighborhood best positions. That is,

$$\lim_{x \to \infty} x_{ij} = \frac{c_1 y_{ij} + c_2 \hat{y}_{ij}}{c_1 + c_2} \dots (3)$$

This theoretically derived behavior provides support for the barebones PSO developed by Kennedy [10], where the velocity vector is replaced with a vector of random numbers sampled from a Gaussian distribution with the mean defined by equation (3), assuming that c_1 =

 c_2 , and deviation,

$$\sigma = \left| y_{ij} - \hat{y}_{ij} \right|$$

The velocity equation changes to

$$v_{ij}(t+1) \sim N(\frac{y_{ij} + \hat{y}_{ij}}{2}, \sigma)$$

The position update then changes to

$$x_i(t+1) = v_i(t+1)$$

Kennedy [9] also proposed an alternative version of the barebones PSO, where

$$v_{ij}(t+1) = \begin{cases} y_{ij} & if \ U(0,1) < 0.5 \\ N(\frac{y_{ij} + \hat{y}_{ij}}{2}, \sigma) & otherwise \end{cases} (4)$$

Based on equation (4), there is a 50% chance that the j-th dimension of the particle dimension changes to the corresponding personal best position. This version of the barebones PSO biases towards exploiting personal best positions.

II. PROBLEM DEFINITION

The basic logical structure of nesting problem investigated in this research paper presumes that the stock sheet/object and items to be nested are rectangular in nature, which categorizes it as Rectangle packing/Nesting. Accordingly, there is a stock of large geometrically defined objects stock sheets that are basically rectangles with different dimensions and an order list for small geometrically defined items (ordered items) also

rectangles of diverse lengths and widths. A layout pattern is defined as an arrangement indicating various shapes that can be laid on a stock sheet. So a solution to a nesting problem consists of a set of layout patterns together with instructions as to how often each layout pattern is to be used so as to meet the requirement.

The problem is solved:-

- A. Using LPP approach in a holistic manner when each stock sheet is associated with different layouts each that guarantee different sheet utilization value and layout pattern as suggested by the rectangle packing heuristic [Sing and Jain, 2009]. The solution is obtained by solving LPP in a holistic manner by ceiling floating-point solutions to integer values.
- B. By exhausting requirement regularly by making use of rectangle packing heuristic by considering the layout pattern that result in best sheet utilization.

Hypothesis: - The results produced by the two approaches do not differ substantially.

III. MATHEMATICS FORMULATION OF LINEAR PROGRAMMING PROBLEM

Let

- K Maximum number of feasible layout patterns considered for each available sheet.
- *J* Number of possible stock sheets available in the reserve. $1 \le j \le m$
- I Number of possible objects in the order list $1 \le i \le n$
- P_{ijk} Possible pieces of the i^{th} space found on the
- j^{th} stock sheet in k^{th} layout pattern.
- X_{ijk} Number of times k^{th} layout pattern of j^{th} stock sheet enter into optimal solution of the problem.
- q_i Requirement for each i^{th} shape as per the order list. $1 \le i \le n$
- l_j Maximum number of j^{th} stock sheet available in the reserve. $1 \le j \le m$
- T_{ik} Trim loss associated with each k^{th} layout

- pattern of j^{th} stock sheet
- A_j Area of each j^{th} stock sheet available in the reserve.
- a_i Area of each i^{th} object in the order list.
- TPC Total number of feasible pattern considered in LPP.

Objective function is defined as:

$$\begin{aligned} & \textit{Maximize} \sum_{j} (\sum_{k} (UF_{jk}))(X_{jk}) & 1 \leq k \leq TPC; \ 1 \leq j \leq m; \\ & \textit{Minimize} \sum_{i} (\sum_{k} T_{jk})(X_{jk}) & 1 \leq k \leq TPC; \ 1 \leq j \leq m; \end{aligned}$$

Subject to the following constraints

 Requirement for each object in the order list

$$\sum_{i} (\sum_{k} (P_{ijk})(X_{jk}) \ge q_i \quad \forall i, \ 1 \le k \le TPC, \ 1 \le j \le m, \ 1 \le i \le n.$$

Reserve of stock sheets

$$\sum\nolimits_{k} X_{ijk} \geq l_{i} \quad \forall j, \ 1 \leq k \leq TPC; \ 1 \leq j \leq m;$$

• Natural Constraint

$$X_{ij} \ge 0$$
 $\forall j, 1 \le k \le TPC; 1 \le j \le m;$

• Feasibility Constraint

$$\sum{}_{i}l_{j}A_{j}\geq\sum q_{i}a_{i}\qquad 1\leq j\leq m;\, 1\leq i\leq n;$$

IV. BRIEF SYNOPSIS OF HEURISTIC

One such heuristic study which is focused to 2-Dim rectangular packing problem where rectangular items are packed on to a larger containing region rectangular in shape, say object. The specific problem addressed is characterized by a set of rectangular items, which may contain identical items that can be rotated by 90 degree during packing process; packing arrangements are non-guillotine and orthogonal and the stock sheet is rectangular in shape, without any bad patches in it. Also, grain orientation of the item and object are trivial. The packing process has to ensure that there is no overlap of items while confining within the object.

A. The items are first sequenced. The criteria for sequencing is based on the all-possible

aspects related to a rectangle e.g., length, width, area, perimeter, length/width ratio, priority assigned to items. This priority is expressed in terms of urgency for an ordered piece or profit associated with it. The lists are sorted in increasing order and in decreasing order.

B. The item thus picked first is placed horizontally and then vertically at the lower left corner of the object, referred to as its reference point. Also, it is important to define here Sheet Utilization Ratio as the sum total area of ail items placed on the stock sheet to total stock sheet area. The item placed at reference point of the object gives rise to two pivot points. Pivot are the top-left and bottom-right corner of the ordered piece placed and the top-left and bottom-right corner of the ordered piece placed and the top-left and bottom-right corners of enclosing rectangle that encloses all ordered pieces placed so far. Essentially, pivot points are the only probable positions where next item in sequence can be placed. The next item is then placed at each of the pivot points, both length-wise and breadthwise for all feasible results. Orientations that result in minimum wastage are retained.

New pivot points are defined and used ones are deleted. In the algorithm, pivot points are sequenced in the following three ways:-

- Minimum radial distance (PPD)
- Minimum x-distance, in case of tie, point with minimum y-distance is to be served first (PPL)
- Minimum y-distance, in case of tie, point with minimum y-distance is to be served first (PPB)

Total number of feasible layouts thus possible are the product of possible orientations of the object (2) X possible orientations of the first item placed on the object (2) X possible sequencing patterns (increasing and decreasing) (2) X different basis for sequencing of items considered (5) X different sequencings of pivot points (along) considered (3)=2 X 2 X 2 X 5 X 3 = 120; Thus OL-IL-I-SL-PPD stands for pattern obtained when object is oriented length-wise; first-item placed on the reference point is also oriented length-wise; item are sorted in increasing order and are sequenced on the basis of length; pivot points are arranged in increasing order along the Diagonal of the object.

V. EXPERIMENT SETUP

Test data set, consisting of the objects of different sizes and the items required to the placed on these objects, were considered as suggested in Table I i.e., SET-I.

SET-I	STOCK SHEET : 2										
		LENGTH = 70 WIDTH = 40 QUANTITY: 5									
			LEN	GTH = 4	0 WIDTI	H = 40 Q	UANTIT	Y:4			
	REQIREMENT										
SR.No	1	2	3	4	5	6	7	8	9	10	
LENGTH	22	31	35	24	30	13	14	14	12	13	
WIDTH	21	13	9	9	7	11	10	8	8	7	
QUANTITY	7	9	4	9	12	4	11	3	12	14	

VI. RESULTS

The revised heuristic has been used on Set-I to generate 120 feasible patterns on each type of available sheet. In total 34 different layouts

were observed (22 on sheet (70×40)). All these layouts have been tabulated in Table 2 and Table 3.

Table 2 Different patterns generated by sheet type 1 ((70×40))

PIECES	22X2	31X13	35X9	24X9	30X7	13X11	14X10	14X8	12X8	13X7	Utilization
	1										Factor (UF %)
QUANTITY	7	9	4	9	12	4	11	3	12	14	
Pattern 1	0	0	0	0	0	4	0	0	12	7	84.31
Pattern 2	0	0	0	0	0	4	0	0	12	8	87.57
Pattern 3	0	0	0	0	11	0	0	0	0	0	82.5
Pattern 4	0	0	0	0	12	0	0	0	0	0	90
Pattern 5	0	0	0	0	0	0	0	0	12	14	86.64
Pattern 6	0	0	0	0	0	0	0	0	11	13	79.64
Pattern 7	0	0	0	0	0	0	0	1	12	14	90.64
Pattern 8	3	0	0	0	0	4	1	0	5	0	92.07
Pattern 9	3	0	0	0	0	4	2	0	0	1	83.18
Pattern 10	3	0	0	0	0	4	1	0	1	5	94.61
Pattern 11	3	0	0	0	0	4	2	0	0	0	79.93
Pattern 12	0	2	4	0	2	0	0	0	0	0	88.79
Pattern 13	0	2	4	0	2	0	0	1	0	0	92.79
Pattern 14	3	2	0	0	0	0	0	1	0	0	82.29
Pattern 15	3	2	0	0	1	0	0	0	0	0	85.79
Pattern 16	0	6	1	0	0	0	0	0	0	0	97.61
Pattern 17	0	5	2	0	0	0	0	0	0	0	94.46
Pattern 18	0	6	0	0	1	0	0	0	0	0	93.86
Pattern 19	0	0	0	0	0	4	0	0	12	5	77.82
Pattern 20	3	2	0	0	0	0	0	1	0	1	85.54
Pattern 21	0	0	0	0	0	0	0	0	11	14	83.21
Pattern 22	0	0	0	0	0	0	0	0	12	13	83.39

Table 3 Different patterns generated by sheet type 2 (40×40)

PIECES	22X21	31X13	35X9	24X9	30X7	13X11	14X10	14X8	12X8	13X7	Utilization
QUANTITY	7	9	4	9	12	4	11	3	12	14	Factor (UF %)
Pattern 23	0	0	0	0	0	2	0	0	12	0	89.89
Pattern 24	0	0	0	0	0	0	0	0	12	3	89.06
Pattern 25	0	0	0	0	6	0	0	0	0	0	78.75
Pattern 26	0	0	0	0	0	0	0	0	0	14	79.63
Pattern 27	0	0	0	0	0	0	0	0	1	14	85.63
Pattern 28	0	0	0	0	0	0	0	0	12	2	83.38
Pattern 29	1	0	0	0	0	4	0	0	2	2	88
Pattern 30	1	0	0	0	0	4	1	0	1	2	90.75
Pattern 31	0	0	4	0	0	0	0	0	0	0	78.75
Pattern 32	1	1	0	0	0	1	0	2	1	0	83
Pattern 33	1	1	0	0	0	1	0	2	0	1	82.69
Pattern 34	0	3	1	0	0	0	0	0	0	0	95.25

In the corresponding LPP, let x_1 represents number of times patterns 1 is to be used, x_2 represents number of times pattern 2 is to be used and so on. Accordingly the mathematical formulation for Set-I is as suggested below and obtained result is summarized in table 4.

Objective Function

 $\begin{aligned} & \text{Maximize} \ z = & 84.31 x_1 + 87.57 x_2 + 90 x_3 + 86.64 x_4 + 90.64 x_5 + 92.07 x_6 + 83.18 x_7 + 94.61 x_8 + 79.93 x_9 \\ & + 88.79 x_{10} + 92.79 x_{11} + 82.29 x_{12} + 85.79 x_{13} + 97.61 x_{14} + 94.46 x_{15} + 93.86 x_{16} + 77.82 x_{17} + 85.54 x_{18} + \\ & 82.5 x_{19} + 79.96 x_{20} + 83.21 x_{21} + 83.39 x_{22} + 89.89 x_{22} + 89.06 x_{24} + 78.75 x_{25} + 79.63 x_{26} + 85.63 x_{27} + 83.38 x_{28} \\ & + 88 x_{29} + 90.75 x_{30} + 78.75 x_{31} + 83 x_{32} + 82.69 x_{33} + 95.25 x_{34} \end{aligned}$

Subjective to Constraints (Requirement for each object in the order list)

1.
$$3x_6 + 3x_7 + 3x_8 + 3x_9 + 3x_{12} + 3x_{13} + 3x_{18} + x_{29} + x_{30} + x_{32} + x_{33} = 7$$

2.
$$2x_{10} + 3x_7 + 3x_8 + 3x_9 + 3x_{12} + 3x_{13} + 3x_{18} + x_{28} + x_{30} + x_{32} + x_{33} + 3x_{34} = 9$$

3.
$$4x_{10} + 4x_{11} + x_{14} + 2x_{15} + 4x_{31} + x_{34} = 4$$

4.
$$12x_3 + 2x_{10} + 2x_{11} + x_{13} + x_{16} + 11x_{19} + 6x_{25} = 12$$

5.
$$4x_1 + 4x_2 + 4x_6 + 4x_7 + 4x_8 + 4x_9 + 4x_{17} + 2x_{23} + 4x_{29} + 4x_{30} + x_{32} + x_{33} = 4$$

6.
$$x_6 + 2x_7 + x_8 + 2x_9 + x_{30} = 11$$

7.
$$x_5 + 2x_{11} + x_{12} + x_{18} + 2x_{32} + 2x_{33} = 3$$

$$8 \ 12x_1 + 12x_2 + 12x_4 + 12x_5 + 5x_6 + x_8 + 12x_{17} + 11x_{20} + 11x_{21} + 12x_{22} + 12x_{23} + 12x_{24} + x_{27}$$

$$+12x_{28} +2x_{29} +x_{30} +x_{32} = 12$$

$$9. \quad 7x_{1} + 8x_{2} + 14x_{4} + 14x_{5} + x_{7} + 5x_{8} + 5x_{17} + x_{18} + 13x_{20} + 14x_{21} + 13x_{22} + 3x_{24} + 14x_{26} \\ + 14x_{27} + 2x_{28} + 2x_{29} + 2x_{30} + x_{33} - 14$$

10.
$$x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10} + x_{11} + x_{12} + x_{13} + x_{14} + x_{15} + x_{16} + x_{17} + x_{18} + x_{19} + x_{20} + x_{21} + x_{22} = 5$$

11. $x_{23} + x_{24} + x_{25} + x_{26} + x_{27} + x_{28} + x_{29} + x_{30} + x_{31} + x_{32} + x_{33} + x_{34} = 9$

[Natural Constraints]

$$x_i \ge 0$$
 $i = 1, 2, \dots, 34$

Optimal Solution of the Rectangle Packing Problem is:

$$x_1 = x_2 = x_4 = x_5 = x_6 = x_8 = x_9 = x_{10} = x_{12} = x_{13} = x_{15}$$

$$= x_{16} = x_{17} = x_{18} = x_{19} = x_{20} = x_{21} = x_{22} = 0$$

$$x_{23} = x_{24} = x_{25} = x_{26} = x_{27} = x_{28} = x_{29} = x_{31} = x_{33} = x_{34} = 0$$

$$x_3 = x_6 = x_7 = x_{11} = x_{14} = x_{32} = 1$$
; $x_{30} = 8$

[Reserve of stock sheets]

Table: 4 Description of result produced by LPP of SET-I

Variable	TPP	UF (%)	No. of Sheet	Sheet Type	Heuristic
X ₃	12	90	1	1	OB-IL-I-SW-PPL
X ₆	27	90.64	1	1	OL-IB-I-SA-PPL
X ₇	10	83.18	1	1	OB-IL-I-SAR-PPL
X ₁₁	9	92.79	1	1	OB-IB-D-SL-PPL
X ₁₄	7	93.86	1	1	OB-IB-D-SL-PPL
X ₃₀	9	90.75	8	2	OB-IB-I-SAR-PPL
X ₃₂	6	83	1	2	OB-IB-D-SB-PPL

Now the revised heuristic is applied in steps while selecting the best layout (A solution with maximum sheet utilization factor) and step by step exhausting the demand for different pieces.

The solution obtained is tabulated (table 5) below.

Table: 5 Description of result produced by Revised Rectangle Packing Heuristic of SET-I

UF (%)	TPP	No. of Sheet	Sheet Type	Heuristic
97.61	07	1	1	OL-IL-D-SP-PPL
95.35	09	1	1	OB-IB-D-SP-PPL
96.21	10	1	1	OL-IB-D-SA-PPL
97.68	15	1	1	OL-IB-D-SL-PPL
95.82	14	1	1	OL-IB-D-SB-PPL
89.94	09	1	2	OL-IL-D-SL-PPL
89.44	15	1	2	OL-IL-D-SAR-PPL
67.13	06	1	2	OL-IB-D-SL-PPL

VII. CONCLUSION

On comparing the two results as summarized in Table 4 and Table 5 for Set-I, it has been observed that solution obtained by using revised heuristic in

and step by step exhausting the demand in comparison to LPP meeting demand wholesomely makes use of smaller number of stock sheets and produces lesser number of surplus ordered pieces.

Table: 6 Comparing between Linear Programming Packing Solution and Revised Rectangle Packing Heuristic Packing Solution for Set-I

I	Revised Rectangle Heuristic Packing Solution	Linear Programming Packing Solution	
	085	143	Total pieces placed
	005	005	Sheets (70x40) used
	003	009	Sheets (40x40) used
	008	014	Total Sheets used in the approach
	000	058	Surplus pieces

REFERENCES

- [1] R.C. Eberhart and J. Kennedy. A New Optimizer using Particle Swarm Theory. In Proceedings of the Sixth International Symposium on Micromachine and Human Science, pages 39–43, 1995.
- [2] J. Kennedy and R.C. Eberhart. Particle Swarm Optimization. In Proceedings of the IEEE International Joint Conference on Neural Networks, pages 1942–1948. IEEE Press, 1995.
- [3] J. Kennedy. Small Worlds and Mega-Minds: Effects of Neighborhood Topology on Particle Swarm Performance. In Proceedings of the IEEE Congress on Evolutionary Computation, volume 3, pages 1931– 1938, July 1999.
- [4] J. Kennedy and R. Mendes. Population Structure and Particle Performance. In Proceedings of the IEEE Congress on Evolutionary Computation, pages 1671–1676. IEEE Press, 2002.
- [5] E.S. Peer, F. van den Bergh, and A.P. Engelbrecht. Using Neighborhoods with the Guaranteed Convergence PSO. In Proceedings of the IEEE Swarm Intelligence Symposium, pages 235–242. IEEE Press, 2003.
- [6] A.P. Engelbrecht. Fundamentals of Computational Swarm Intelligence. Wiley & Sons, 2005.

- [7] J. Kennedy, R.C. Eberhart, and Y. Shi. Swarm Intelligence. Morgan Kaufmann, 2001.
- [8] F. van den Bergh. An Analysis of Particle Swarm Optimizers. PhD thesis, Department of Computer Science, University of Pretoria, Pretoria, South Africa, 2002.
- [9] F. van den Bergh and A.P. Engelbrecht. A Study of Particle Swarm Optimization Particle Trajectories. Information Sciences, 176(8):937–971, 2006.
- [10] J. Kennedy. Bare Bones Particle Swarms. In Proceedings of the IEEE Swarm Intelligence Symposium, pages 80–87, April 2003.
- [11] Y. Shi and R.C. Eberhart. A Modified Particle Swarm Optimizer. In Proceedings of the IEEE Congress on Evolutionary Computation, pages 69–73, May 1998.

Localization Accuracy Improved Methods Based on Adaptive Weighted Centroid Localization Algorithm in Wireless Sensor Networks

Chang-Woo Song, Jun-Ling Ma, Jung-Hyun Lee Department of Information Engineering, INHA University, Incheon, Korea,

Kyung-Yong Chung Department of Computer Information Engineering, Sangji University, Wonju, Korea Kee-Wook Rim
Department of Computer and
Information Science, Sunmoon
University, Asan, Korea

Abstract—Generally, see Localization of nodes is a key technology for application of wireless sensor network. Having a GPS receiver on every sensor node is costly. In the past, several approaches, including range-based and range-free, have been proposed to calculate positions for randomly deployed sensor nodes. Most of them use some special nodes, called anchor nodes, which are assumed to know their own locations. Other sensors compute their locations based on the information provided by these anchor nodes. This paper uses a single mobile anchor node to move in the sensing field and broadcast its current position periodically. We provide an adaptive weighted centroid localization algorithm that uses coefficients, which are decided by the influence of mobile anchor node to unknown nodes, to prompt localization accuracy. We also suggest a criterion which is used to select mobile anchor node which involve in computing the position of nodes for improving localization accuracy. The localization accuracy of adaptive weighted centroid localization algorithm is better than maximum likelihood estimation which is used very often.

Keywords-component; Weighted Centroid Algorithm; Wireless Sensor Networks; Localization;

I. INTRODUCTION

A wireless sensor network (WSN) consists of spatially distributed autonomous sensors to cooperatively monitor physical or environmental conditions, such as temperature, sound, vibration, pressure, motion or pollutants. The development of wireless sensor networks was motivated by military applications such as battlefield surveillance. They are now used in many industrial and civilian application areas, including industrial process monitoring and control, machine health monitoring, environment and habitat monitoring, healthcare applications, home automation, and traffic control.

A sensor network normally constitutes a wireless ad-hoc network, meaning that each sensor supports a multi-hop routing algorithm (several nodes may forward data packets to the base station). In computer science and telecommunications, wireless sensor networks are an active research area with numerous workshops and conferences arranged each year. The applications for WSNs are varied, typically involving some

kind of monitoring, tracking, or controlling. Specific applications include habitat monitoring, object tracking, nuclear reactor control, fire detection, and traffic monitoring. In a typical application, a WSN is scattered in a region where it is meant to collect data through its sensor nodes.

A sensor network is composed of a large number of sensor nodes that are densely deployed in a field. Each sensor performs a sensing task for detecting specific events. The sink, which is a particular node, is responsible for collecting sensing data reported from all the sensors, and finally transmits the data to a task manager. If the sensors can't directly communicate with the sink, some intermediate sensors have to forward the data [1].

There are several essential issues (e.g., localization, deployment, and coverage) in wireless sensor networks. Localization is one of the most important subjects for wireless sensor networks since many applications such as environment monitoring, vehicle tracking and mapping depend on knowing the locations of the sensor nodes. In addition, with location-based routing protocols, both routing and data forwarding are determined based on the geographic location [2].

To solve the localization problem, it is natural to consider placing sensors manually or equipping each sensor with a GPS receiver. However, due to the large scale nature of sensor networks, those two methods become either inefficient or costly, so researchers propose to use a variety of localization approaches for sensor network localization.

These approaches can be classified as range-based and range-free. Firstly, the range-based approach uses an absolute node-to-node distance or angle between neighboring sensors to estimate locations. Common techniques for distance or angle estimation include received signal strength indicator (RSSI), time of arrival (TOA), time difference of arrival (TDOA), and angle of arrival (AOA). The approaches typically have higher location accuracy but require additional hardware to measure distances or angles. Secondly, the range-free approach does not need the distance or angle information for localization, and depends only on connectivity of the network and the contents

Vol. 8, No. 8, November 2010

of received messages. For example, Centroid method, APIT method, DV-HOP method, Convex hull, Bounding box, and Amorphous algorithm have been proposed [3][4][5]. Although the range-free approach cannot accomplish as high precision as the range-based [6], they provide an economic approach. Due to the inherent characteristics (low power and cost) of wireless sensor networks, the range-free mechanism could be a better choice to localize a sensor's position, so we pay more attention to range-free approach in this paper.

This paper uses a single mobile anchor node as the reference node, which is required to move in the sensing field and broadcast its current position periodically. Sensor nodes receive the position information of the mobile node and localize themselves to the centroid of these positions by using adaptive weighted centroid algorithm. The algorithm based on the Received Signal Strength Indication (RSSI). The results of simulations show that the method is a practical method that can be used in real-world system, and is also a method whose principle is simple, less computing and communication, is low cost, and provides flexible accuracy.

RELATED WORK

In the past several years, extensive research has been done on localization for wireless sensor networks. A general survey is found in. Here we provide only a brief survey about rangefree approaches and localization method, which involve mobile reference nodes. Some nodes are equipped with special positioning devices that are aware of their locations. These nodes are called anchor nodes or reference nodes. Other nodes that do not initially know their locations are called unknown nodes or sensor nodes. Generally, an unknown node can estimate its location by range-based or range-free methods if three or more anchors are available in its coverage field. Obviously, the number and position of anchor nodes have a noticeable influence on the localization precision.

The main idea of localization with a mobile anchor node is as follows: After sensor deployment, a mobile anchor node traverses the sensor network while broadcasting anchor packets, which contain the coordinates of the anchor node. Sensor nodes receiving anchor packets could infer their distance from a mobile anchor node and use these measurements as constraints to construct and maintain position estimates. These methods have a common feature: they use range-based approaches. Though they can reach fine resolution, either the required hardware is expensive (ultrasound devices for TDOA, antenna arrays for AOA) or the results depend on other unrealistic assumptions about signal propagation (for example, the actual received signal strengths of radio signals can vary when the surrounding environment changes). Due to the hardware limitations of sensor devices, range-free approaches are a cost effective alternative to a more expensive range-based approach. A simple algorithm proposed, computes location as the centroid of its proximate anchor nodes. An alternate solution, DV-Hop, extends the single hop broadcast to multiple-hop flooding, so that sensors can find their distance from the anchors in terms of hop counts. An amorphous positioning scheme adopts a similar strategy as DV-Hop; the major difference is that Amorphous improves location

estimates using offline hop-distance estimations through neighbor information exchange. Another existing range-free scheme is an APIT algorithm. APIT resolves the localization problem by isolating the environment into triangular regions between anchor nodes. A node uses the point-in-triangle test to determine its relative location with triangles formed by anchors and thus narrows down the area in which it probably resides. APIT defines the center of gravity of the intersection of all triangles that a node resides in as the estimated node location [7][8][9].

Based on these analyses, localization using a single mobile anchor node would be more economical. In addition, considering the constraints in computing and memory power of sensors, we adopted the weighted centroid method with a single mobile anchor to locate sensors in wireless sensor networks. Depending on the method used for ranging, an appropriate localization technique is applied in the second phase. The following localization strategies have been proposed.

A. Trilateration

This is one of the more popular strategies and is used when the exact distances between known points and an object to be located are available. Fig. 1 shows when the distance between an object and three points are given, the object's location x can be computed as the intersection of three circles centered at the known points.

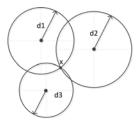


Figure 1. Example of Trilateration

B. Bounded Intersection

The trilateration technique works well when the three circles intersect at a single point, but this is rarely the case when estimates are used in ranging. For example, when using incremental stepping of transmission power for ranging, maximum values can be used for estimating the distances. Fig. 2 shows The object to be located would fall into a geometric region that is the intersection of three circles.

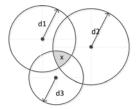


Figure 2. Localization with Maximum Bounds

C. Triangulation

The triangulation method is useful if the angle between two objects can be measured. Fig. 3 shows an example. Suppose P1 and P2 are points with known locations and X is an object to be located. Nodes P1 and P2 can measure angles a1 and a2 , and, with known distance Sx, one can easily compute ax, S1 and S2.

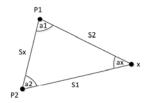


Figure 3. Example of Triangulation

D. Maximum Likelihood

When estimates are used for ranging, it is possible that the region of intersection is empty. This will occur if at least one ranging estimate is too small. One method that overcomes this problem selects the point for localization that gives the minimum total error between measured estimates and distances. In Fig. 4, distance estimates (d1, d2, d3) are made between the object to be located and three points (P1, P2, P3). The errors (e1, e2, e3) are computed by finding the difference between the actual Euclidean distances and the ranging estimates.

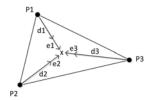


Figure 4. Localization with Maximum Likelihood

III. ADAPTIVE WEIGHTED CENTROID LOCALIZATION METHOD

A. Method of Localization

This method can be used in large-scale field environment. Figure 1 illustrates the system environment where a sensor network consists of a mobile anchor node and unknown nodes that could be scattered from a plane or from a mortar shell. The mobile anchor is a human operator or an unmanned vehicle deployed with the sensor network. If the network is deployed by plane scattering, this anchor can be even the plane itself. The unknown nodes are the nodes of initially unknown positions. Once the nodes are deployed, they will stay at their locations to conduct the sensing task. The mobile anchor, which is a node aware of its location (e.g. equipped with GPS), and is able to traverse for assisting the sensors to determine other node locations [10]. The mobile anchor node needs to traverse over the entire region in order to cover all sensor

nodes. This can be done by driving the mobile anchor node to move in a spiral trajectory. Obviously there are many other options to moving trajectory. Finding an optimal trajectory to cover all sensor nodes can be a research topic for our future work. No matter which trajectory is used, the location of the mobile anchor node on the trajectory should be known. At the same time, we assume that the mobile anchor has sufficient energy for moving and broadcasting its information during the localization process. The speed of the mobile anchor is adjustable and unrestricted, but uniform in the process of location.

We used an idealized radio model for wireless communication because it was simple and easy to reason mathematically. We assumed that our idealized model is perfect spherical radio propagation and has identical transmission range (power) for all radio positions as shown in Fig. 5. It is a sphere with the anchor as its center and the broadcasting radius R as its radius. Only the sensors within the range are assumed capable of receiving the information sent by the anchor.

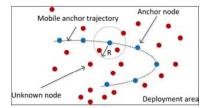


Figure 5. System Environment with a Mobile Anchor Node

In this paper, we proposed the location of mobile anchor node influence: In the localization algorithm, location of mobile anchor node has influence to the unknown nodes, RSSI bigger location, and the greater influence on the location of sensor nodes. When Unknown node received multiple mobile anchor node position signal then unknown node by the impact of these locations. Location of largest RSSI has the greatest power to decide to the position of sensor node.

Signal selection principle: An unknown node may receive multiple signals of positions from the mobile anchor node.

RSSI value should be the largest of several signals position calculation. Location computed to ensure that the signals involved in more than three. Will be distances of more than R the location of mobile anchor node removed, so as to avoid the expansion of the positioning error. Behind the simulation proves this point.

B. Adaptive Weighted Centroid Localization Algorithm

Through the front of the Analysis, can find common centroid algorithm, did not reflect the mobile anchor node's influence, affecting the localization accuracy. To enhance the localization accuracy, in this paper we used the adaptive weighted centroid localization algorithm. Its main idea: In the algorithm, mobile anchor node confronts the right to decide the location of the centroid through weighted factor to reflect. The use of weighted factor reflected the intrinsic relationship between them.

We embody this relationship through the formula of the weighted factor:

$$X = \frac{\left(\frac{X1}{d1} + \frac{X2}{d2} + \frac{X3}{d3}\right)}{\left(\frac{1}{d1} + \frac{1}{d2} + \frac{1}{d3}\right)}, Y = \frac{\left(\frac{Y1}{d1} + \frac{Y2}{d2} + \frac{Y3}{d3}\right)}{\left(\frac{1}{d1} + \frac{1}{d2} + \frac{1}{d3}\right)} \quad (1)$$

Fig. 6 illustrates Known 3 mobile anchor nodes coordinate (X1, Y1), (X2, Y2), (X3, Y3), unknown node to anchor nodes distance d1, d2, d3. According to the formula can be calculated unknown node coordinates (X, Y). Compared to ordinary centroid algorithm, 1/ d1, 1/ d2, 1/ d3 is weighted factor. The factor 1/ d1, 1/ d2, 1/ d3 indicates that mobile anchor node with a shorter distance to unknown nodes has a larger infect its coordinates. We can improve the localization accuracy from these inner relations.

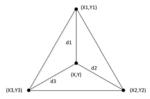


Figure 6. Scheme of the adaptive weighted centroid localization algorithm

Adaptive Weighted Centroid Localization Algorithm process:

- 1) The mobile anchor node periodically sends its own information.
- 2) Unknown node received information, only records the same location of the mobile anchor node average RSSI.
- 3) Unknown node received over threshold m in the position information then RSSI value in accordance with the smallest sort of mobile anchor node location. And to establish the mapping between RSSI value and the distance from unknown node to the mobile anchor node. The establishment of three sets: mobile anchor node_set={a1, a2, ..., am}; Distance_set={d1, d2, ..., d m}; Mobile anchor node position_set={(X1, Y1), (X2, Y2), ..., (Xm, Ym)};
- 4) RSSI value with the first few large location of mobile anchor node of the calculation:
- a) Based on the preceding analysis, In the mobile anchor node_set Select RSSI value of large node location then the composition of the triangle set. This is very important. Triangle_set={(a1, a2, a3), (a1, a2, a4), ...(a1, a3, a4), (a1, a3, a5)...};
- 5) n location of mobile anchor nodes can be composed of C_n^3 triangles. The use formula (1) calculates C_n^3 coordinate.

6) Calculates the mean value(X,Y) of C_n^3 coordinate. The (X,Y) is Unknown node coordinate.

IV. EXPERIMENTS

A. Simulation Environment

The key metric for evaluating a localization technique is the accuracy of the location estimates versus the communication and deployment cost. To evaluate this proposed method we use UNIX, programs with the C language. We have carried on the computer simulation to the above algorithm. Simulation condition: The mobile anchor node reference MICA2 mote; Uses outdoor launches the radius 200 to 300m; Deployment area is 200*200. The unknown node arranges stochastically; the unknown node is 220. The mobile anchor node has 6 kinds of situations: 9, 12, 16, 20, 25, and 30 positions.

B. Results and Analysis

The simulation uses adaptive weighted centroid localization algorithm and maximum likelihood estimation method. Localization accuracy mainly depends on the numbers of the mobile anchor node broadcasting its positions or the anchor density. It is very easy for our method to change anchor density by adjusting the interval time or the moving length of the mobile anchor node broadcasting its positions or by changing the moving interval of spiral line. In comparison with other methods, this is one of the advantages with our method, and it does not require additional hardware. Figure 7 and Figure 8 show the simulation result. In the figure 7 error of adaptive weighted centroid localization algorithm is 16.2m and error of maximum likelihood estimation is 24.2m when mobile anchor node is 9. In the figure 8 error of adaptive weighted centroid localization algorithm is 21.5m and error of maximum likelihood estimation is 40.5m when mobile anchor node is 30. As can be seen from the figure adaptive weighted centroid localization algorithm has better localization accuracy. Has the obvious superiority, if the anchor density is low. Adaptive Weighted Centroid Localization Algorithm is simple, and no communication is needed while locating. It does not require additional hardware. The mobile anchor node can be used many times. So it is very inexpensive.

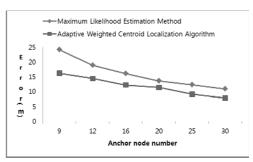


Figure 7. Average Error

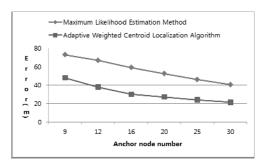


Figure 8. Maximum Error

V. CONCLUSION AND FUTURE WORKS

The Many wireless sensor network applications depend on nodes being able to accurately determine their locations. This is the first work to study range-free localization in the presence of mobility. One of our ideas is that a mobile anchor can improve the localization accuracy and coverage because it can move to every point of wireless sensor networks. Another factor is that range-free requires no extra hardware or data communication and reduces the costs of localization. Our simulation experiments reveal that our method can provide accurate localization even when memory limits are severe, the seed density is low, and network transmissions are highly irregular.

Many issues remain to be explored in future work including how to select a moving path to improve the locating performance, how to apply this to real-world sensor networks and how to expend our method to other applications.

ACKNOWLEDGMENT

"This research was supported by the MKE(The Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) Support program supervised by the NIPA(National IT industry Promotion Agency)" (NIPA-2010-C1090-1031-0004).

REFERENCES

- Ssu, K. F., Ou, C. H., and Jiau, H. C.:Localization with mobile anchor points in wireless sensor networks. IEEE Trans. on Vehicular Technology, Vol. 54, No. 3, pp. 1187-1197, 2005.
- [2] Hu, L., and Evans, D.:Localization for mobile sensor networks. in Proc. of ACM MobiCom, 2004.
- [3] He, T., Huang, C., Blum, B. M., Stankovic, J. A., and Abdelzaher, T.:Range-free localization schemes for large scale sensor networks. in Proc. ACM Int. Conf. Mobile Computing Networking (MOBICOM), San Diego, CA, pp. 81-95, 2003.
- [4] Niculescu, D., and Nath, B.:DV based positioning in ad hoc networks. Kluwer J. Telecommun. Syst, Vol. 22, No. 1, pp. 267-280, 2003.

- [5] Nagpal, R., Shrobe, H., and Bachrach, J.:Organizing a global coordinate system from local information on an ad hoc sensor network. In IPSN'03, 2003
- [6] Niculescu, D., Nath, B.:DV based positioning in ad hoc networks. Journal of Telecommunication Systems, Vol. 22, No. 4, pp.267-280, 2002
- [7] Kim, Y. C., Kim, Y. J., Chang, J. W.:Distributed Grid Scheme using S-GRID for Location Information Management of a Large Number of Moving Objects. Journal of Korea Spatial Information System Society, Vol. 10, No. 4, pp.11-19, 2008.
- [8] Lee, Y. K., Jung, Y. J., Ryu, K. H.:Design and Implementation of a System for Environmental Monitoring Sensor Network. In Proc. Conf. APWeb/WAIM Workshop on DataBase Management and Application over Networks, pp. 223-228, 2008.
- [9] Hammad, M. A., Aref, W. G., Elmagarmid, A. K.:Stream window join: Tracking moving objects in sensornetwork databases. In SSDBM, 2003.
- [10] Niculescu, D., and Nath, B.:Position and orientation in ad hoc networks. Ad hoc Networks, Vol. 2, No. 2, pp. 133-151, 2002.

AUTHORS PROFILE

Chang-Woo Song received M.S. degree, from the Inha University, in 2007, respectively. He is currently working as a Lecturer in the Department of Computer System Enginnering, affiliated to Inha Technical College. He research interest includes ubiquitous/embedded system, computer architecture, mobile programming.

Jun-Ling Ma received M.S. degree, from the Inha University, in 2010, respectively. He is doing his research in Wireless Sensor Network. His area of interest includes operating systems and object analysis and design.

Professor **Kyung-Yong Chung** received the B.S., M.S. and Ph.D. degrees from the Inha University, Korea, in 2000, 2002, and 2005, respectively. Currently, he is a Professor in the School of Computer Information Engineering, Sangji University, Korea. His research interest includes data mining, HCI, information retrieval, and sensibility engineering.

Professor Jung-Hyun Lee received the B.S., M.S. and Ph.D. degrees from the Inha University, Korea, in 1977, 1980 and 1988, respectively, all in Electrical Engineering. Since 1989, he is a Professor in the School of Computer Science & Engineering, Inha University. In 1979-1981, he was a researcher at the Korea Institute of Electronics Technology. In 1984-1989, he was an Associate Professor at the Kyonggi University. His research interests are in computer architecture, speech recognition, data mining, HCI, information retrieval, and sensibility engineering.

Professor **Kee-Wook Rim** received the B.S., and Ph.D. degrees, from the Inha University, in 1987 and 1994; the M.S. degree from Hanyang Univercity, respectively. Since 2000, he is a Professor in the School of Computer Science and Engineering, Sunmoon University. In 1977-1999, he was a Senior Researcher at ETRI and TICOM Development Manager. His research interests are in Real-time Database Systems, Operating Systems and System.

A Novel Hybridization of ABC with CBR for Pseudoknotted RNA Structure

Ra'ed M. Al-Khatib, Nur'Aini Abdul Rashid and Rosni Abdullah

School of Computer Science Universiti Sains Malaysia USM Penang, Malaysia

Abstract— The RNA molecule is substantiated to play important functions in living cells. The class of RNA with pseudoknots, has essential roles in designing remedies for many virus diseases in therapeutic domain. These various useful functions can be inferred from RNA secondary structure with pseudoknots. Many computational intensive efforts have been emerged with the aim of predicting the pseudoknotted RNA secondary structure. The computational approaches are much promising to predict the RNA structure. The reason behind this is that, the experimental methods for determining the RNA tertiary structure are difficult, timeconsuming and tedious. In this paper, we introduce ABCRna, a novel method for predicting RNA secondary structure with pseudoknots. method combines heuristic-based This KnotSeeker with a thermodynamic programming model, UNAFold. ABCRna is a hybrid swarm-based intelligence method inspired by the secreting honey process in natural honey-bee colonies. The novel aspect of this method is adapting Case-Based Reasoning (CBR) and knowledge base, two prominent Artificial Intelligence techniques. They are employed particularly to enhance the quality performance of the proposed method. The CBR provides an intelligent decision, which results more accurate predicted RNA structure. This modified ABCRna method is tested using different kinds of RNA sequences to prove and compare its efficiency against other pseudoknotted RNA predicted methods in the literature. The proposed ABCRna algorithm performs faster with significant improvement in accuracy, even for long RNA sequences.

Keywords-RNA secondary structure; pseudoknots; Case-Bases Reasoning; Artificial Bee Colony (ABC) algorithm.

I. INTRODUCTION

Ribonucleic acid or (RNA) is one of the nucleic acids, which plays diverse roles and functions. Basically, one kind of RNA is the messenger RNA (mRNA). It works as an intermediary in carrying the genetic information code from DNA to make proteins [1]. This carried genetic code is used in the natural process for synthesizing proteins in living cell. However, the recent biological studies confirmed that there are other kinds of RNAs, which play various useful roles [2]. The latest discovered functions of RNA molecule, include: splicing introns, catalyst for reaction and a regular in cellular

activities [3, 4]. Predicting the RNA structure is the key to determine and scrutinize the active functions of RNA molecule. This fact is emphasized by central dogma in biochemistry and biology research domain [5, 6]. The RNA secondary structural outputs provide the base for shaping the RNA three-dimension (3D) structure, which is the first step of the RNA tertiary structure phase.

The importance of the computational methods for predicting RNA secondary structure has been acknowledged as a demanding research area, by computer scientists. Also, there are many conditions, facing the experimental methods that are used by biologists [7, 8]. The Nuclear Magnetic Resonance (NMR) and X-ray crystallography are the two popular experimental purification methods that are used to determine the RNA 3D spatial structure [9, 10]. Latest studies confirmed that many classes of RNA molecule broadly fold in the pseudoknot motif [11, 12]. Whereas, the RNA structural functions of pseudoknot elements, have been emphasized to be prominent for medical processes and designing anti-viral treatments, in therapeutic research [13]. Consequently, the computational RNA prediction methods for predicting the RNA secondary structures are extensively utilized with manageable efforts [14].

The RNA molecules come in two main shapes: the Stemloop and the Pseudoknots, as illustrated in Figure 1 in terms of RNA structure classifiers [15]. The Stem-loop is a noncrossing RNA structure motif. While, the Pseudoknots is a crossing RNA structure, which plausibly has been spotted by [16]. Further, the pseudoknotted RNAs has been proven to play several vital roles. From complexity points of view, the top prediction methods of RNA without pseudoknots functional element are MFold [17] and Vienna [18] algorithms which execute with complexities $O(n^3)$ in time and $O(n^2)$ in space. PknotsRG [19] is one of the most proper algorithm for predicting RNA with pseudoknots. It requires $O(n^4)$ and $O(n^2)$ in time and space complexities, respectively. Even if the pseudoknotted RNA secondary structure prediction problem has been stated as Non-deterministic Polynomial time (NP)-Complete problem [20, 21], it is an insisted matter to be solved [22, 23], in recent years.

In order to overcome the prediction problem of RNA secondary structure with pseudoknots, this article introduces

School of Computer Sciences, University Sains Malaysia Penang, Malaysia.

a nature-inspired hybrid method called "ABCRna". Innovatively, this approach combines a new derivation from Artificial Bee Colony (ABC) algorithm with a special deterministic constraints [24]. On top of this, it is borrowed from the Artificial Intelligence (AI) field, which is a kind of nature swarm-intelligence [25]. The objective of this proposed method is to build the entire RNA secondary structure with pseudoknots from a given single-stranded RNA primary sequence. Indeed, this proposed method is a combination of KnotSeeker (heuristic-based method [3]) with UNAFold (a dynamic programming method [26]) for solving the RNA structural related issue. This hybrid method is a new derivation from ABC algorithm. It adapts the inspired swarm-based intelligence behavior of the honeybees in collecting nectar and converting that to honey and royal jelly [27]. Naturally, every individual worker bee visits many flower patches during the round-trip of collecting nectar and pollen. Then it goes back to the hive to submit the mixed nectar to the nurse bee. Finally, the nurse bee starts making honey by a natural biological secreting process.

Intuitively, the proposed RNA structural hybrid method is deployed and built to solve the related pseudoknotted RNA bioinformatics problem. By a deeper understanding of the CBR technique [28], the proposed hybrid model obtains a global optima RNA structural assurance results with more accuracy and better performance. Finally, the results show that the ABCRna method significantly improves the execution time and the accuracy in both sensitivity and specificity. This improvement when comparing the outputs with the other pseudoknotted RNA prediction methods existing in the state-of-the-art like; FlexStem [29], HotKnots [30] and PknotsRG [19].

The remainder of this article is ordered as follows: In the next section, we start with describing the secondary structure of the RNA molecule, in computer context representation. In section 3 background materials, gives a concise expression to the generic ABC optimization method. Then, a derivation of

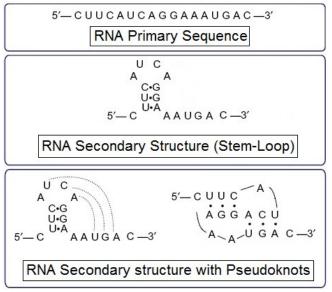


Figure 1. A stem-loop and pseudoknots of RNA structures types.

ABC is adapted to generate the proposed method. Next, the CBR as a modern AI technique, is extensively and widely discussed, from theoretical concept. Section 4 presents the proposed method with the implemental mapping between pseudoknotted RNA secondary structural prediction and the secreting process of making honey. Subsequently, the following section reports the comparative benchmark of the proposed method. The results of ABCRna is comparing against the results of other RNA prediction methods in the literature. Finally, the article ends with conclusion remarks, in section 6.

II. SECONDARY STRUCTURES OF RNA

A. RNA Stem-Loop (non-pseudoknots)

The single-stranded RNA molecule forms many folded structures in hierarchal shape; the primary RNA single sequence, the secondary structure of RNA molecule, the three-dimensional (3D) or tertiary RNA functional structure and the quaternary structure for RNA polymerase [31]. Generally, the RNA computational methods predict the secondary structure of the given RNA primary sequence. Thus, the RNA secondary structure defines: as an RNA structural motif, which in some parts includes the doublestranded motifs. These parts joined by complementary and canonical base pairings with the other parts, which are the non-paired single bases. The double-stranded motif parts coming in several shaped of stem-loops: hairpin, internal (or interior), bulge, multi-branch external bases and stacking (or helices) loops. As explained above and illustrated in Figure 2, the RNA primary sequence (RNA bases) folds and joins on itself in real RNA secondary structure by hydrogen chemical bonds for low energy and more stability [15]. In mathematical and computational representation concept, the various layers of RNA structures can be defined as follows:

- $b = b_1, b_2, ..., b_i, ..., b_n$, where b is an RNA primary sequence and b_i is the RNA base or nucleotide [32, 33]. The element b_i is also a member of set which includes $\{`A', `C', `G', `U', `N'\}$. While, the first four alphabets are representation of the original paired bases (paired-nucleotides) of the real RNA molecule: *Adenine, Cytosine, Guanine* and *Uracil*, respectively. The last nucleotide `N' is assigned to the non-paired base. Such that the n is the length of the given RNA sequence and $1 \le i \le n$.
- $S = \{(b_i, b_j)\}$, such that (b_i, b_j) belongs to the canonical base pairs. S is the secondary structure of the given RNA primary sequence which satisfies the following conditions:
 - (b_i, b_j) ∈ {(A,U), (U,A), (G,C), (C,G), (G,U),(U,G)}, these are the sets of RNA base-pairs. While, the base pairs include in the set {A-U, U-A, G-C, C-G} is a Watson-Crick RNA base-pairs [34], the set {A-U, U-A} is a Wobble RNA base-pair [35].
 - Then $S = \{(b_i, b_j): 1 \le i < j \le n \text{ and } j i > Const \}$, where Const is a threshold constant number depend on the limit length of the minimum un-paired bases in a stem-loop (hairpin, stem or bulge ... etc). The Const is typically taken to be equal three.

- If $(b_i, b_j) \in S$, $(b_k, b_l) \in S$ and if $b_i = b_k$, then $b_j = b_l$. This implies $(b_i, b_j) = (b_k, b_l)$. In another words, every base (nucleotide) in RNA secondary structure make join by hydrogen bond at most with another one base (non-triple or only allow one-to-one).
- If $(b_i, b_j) \in S$, $(b_k, b_l) \in S$ and i < k, this can include two location elements in RNA stem-loop structure (*non-pseudoknots*):
 - If i < k < l < j, then the two base pairs are form a type of nested location elements (nested-fashion), as depictured in Figure 3 a.
 - If i < j < k < l, then the two base pairs are form a type of juxtaposed location elements (juxtaposed-fashion) [36], as shown in Figure 3 b.

B. RNA with Pseudoknots

The majority of RNA molecule classes fold in functional structural elements called pseudoknots. Indeed, they belong to the (3D) tertiary structure element and perform an important useful roles and constructive functions [37].

The pseudoknots substructure can theoretically satisfy the following term. If there are two base pairs (b_i, b_j) and (b_k, b_l) , then satisfy the conditions: i < k < j < l or k < i < l < j, as shown in Figure 3 c and d. These two base paired shapes are represented the pseudoknots RNA structural elements. In another word, the pseudoknots is a crossing sub-structural functional element in the RNA molecules. It forms interaction the unpaired bases part of the stem-loop, which folds back and join in a loop region located outside that stem-loop.

In spite of the prediction algorithms of RNA with pseudoknots structural elements, have been proven to be NP-complete problem [21]. It is a demanding research area because of the pseudoknotted RNAs has importance as key functions. Further it plays essential roles in viral and cellular regulatory [38].

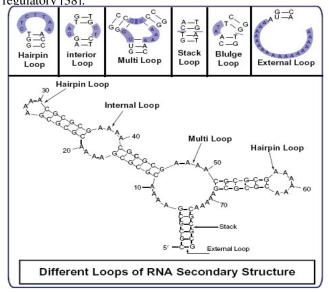


Figure 2. Different RNA element shapes, the image is adapted from [39].

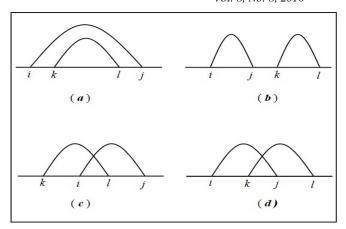


Figure 3. The diagrammatic position relation between different types of RNA base pairs. (a) two base-pair in juxtaposed fashion. (b) two base-pair in nested fashion. (c)&(d) two base-pair in pseudoknots.

III. BACKGROUND MATERIALS

A. Problem Statement of RNA with Pseudoknot

Pseudoknotted RNA secondary structure is the problem of predicting its secondary structure from a given primary sequence. Particularly, it has recently become attractive research area. Due to that the RNA with pseudoknots, has many important and useful roles, which needs to be solved computationally [40]. The existing pseudoknotted RNA prediction algorithms perform in exponential time complexity. The best prediction method run, in the worst case, $O(n^4)$ in time and $O(n^2)$ in space [19]. Thus they run very slowly and need an ever increasing memory-space, especially for long sequences. Veritably, this means that the prediction solving algorithms of the pseudoknotted RNA secondary structural problem, suffer from long execution time and storage complexities. To the best knowledge of the authors, the final structural results suffer from poor quality and inaccuracy, for long RNA sequences.

The pseudoknots class of the RNA structural prediction issue, has been proven an NP-complete problem [20]. Increasingly, the collecting nectar to make honey is an inspired field for the bioinformatics researchers, which is derived from the original ABC model [24]. In this article, a new hybrid method as a sub-area of swarm intelligence approaches for solving the pseudoknotted RNA structural problem is adapted. Besides that the CBR as a modern AI technique highlighted a way to be deployed, in term of enhancement the final results of the proposed hybrid ABCRna model. From comparison points of view, we find this method improved the accuracy of the RNA structural outputs with good performance.

B. Swarm-Intelligence in AI Technology

Swarm Intelligence (SI): is an emergent and bioinspired field of AI, which has been generated from numerous researches in social insect's behavioural models [41]. The phrase swarm comes up to present solution to overcome the optimization problems. These optima solutions have been successfully got by utilizing the co-operative and

Vol. 8, No. 8, 2010

coordinative efforts among the worker-insects. The inspiration of the swarm intelligence is gained from many social insects behavioral models like; honey-bees colony and ant-colony. For instance in bee-colony, the objective of the *swarm* is the quantity and quality production of honey by the mutual teamwork. It is a key fact that, the amount of honey that an individual worker-bee harvests is worthless. But, the honey production by all worker-bees is considerably much better than the crop of an individual one [42].

Lately, swarm intelligence has obtained high interest to be adapted by many researchers from diverse fields. The list compromises, but it is not limited of: engineering, science and commerce fields. The computer researchers propose swarm intelligence optimization methods to solve many complex problems that suffer from severe drawbacks. The typical research domain of the computational swarm intelligence is to solve many real-world problems. Some applications of swarm intelligence in a development areas as follows: (i) The routing optimization in different communication network [43]. (ii) The job scheduling [44]. (iii) The swarm control in the Unmanned Aerial Vehicles (UAV) for both civil-military purposes [45, 46].

C. Honey-Bee Colony Structure

Many social insects live in colonies have instinctual ability to perform as agents in a group for solving complex problems and to complete their tasks. The new AI disciplinary "swarm-intelligence" has been attractively produced by deep knowledge of the biological swarm in solving the problems. This can done by a behavioral interaction among thousands members of the swarm-insects [47]. Naturally, the social insects have talent to be in self-organized behavioral models for achieving an intelligence solution of the vital tasks.

Honey-bees live in a well structured social insect's colony called a hive. The hive typically is a composition of a solo queen, drones and workers [48]. Each one does the following roles: (i) As usual, there is one queen. She is egglaying, female as a mother for other colony members and mates one time in her lifelong by drones. (ii) There are drones or male bees as bee-colony fathers. Their main responsibility is fertilizing the new queen in a mating flight party (social gathering) before dying. They live at most six months and reach to hundreds up to several thousands during the summer season. (iii) There are around 10,000 in winter to 60,000 in summer female worker-bees (foragers) in each bee-colony. They do many important jobs including: collecting nectar to make food, raising and bringing up the broods and larvae's, guarding and ventilating the hive. But, the primary resourceful task of the worker-bees is collecting the nectars and pollens from the flower patches (forage field). Later, when they back to the hive the worker bees secret the honey and royal jelly (food).

D. Honey-bee Collecting Nectar (Foraging)

Honey-bees collecting nectar process to make honey is to be considered as an optimization swarm-based intelligence approach [49]. The worker-bees perform the collecting nectar and secreting honey process in a well-organized behavioral model known as bees foraging process [50]. It is obvious that, this gigantic task is beyond the ability of every worker-bee individually. Nevertheless, all the group members interact among each other in a fashion to solve the collective bee-foraging problem.

The main incentive task in bee's colony is the foraging (collecting nectar to make honey). To investigate the bee foraging process Seeley in [51], introduced a detailed systematic mechanism. It is about the self organized honeybee's social behavioral model in collecting forage, as shown in Figure 4. In the proposed system, every worker bee (forager) visits many flowers from the same type within 30 to 120 minutes of foraging trip. All the collected nectars, from these flower patches, have been stored in the forager honey stomach. Besides that, the forager commits several actions to provide a feedback. Waggle dance is providing the profitability rating of nectar in the flower patches, the odor, location and other required information [52, Accordingly, the making honey and royal jelly process starts when the worker-bee back to hive from the foraging roundtrip journey.

Soon after reaching the hive from the foraging trip, the field bee (forager) gears up to submit that nectar, which already stored in her honey sac [54]. This process of submission the gathered nectar to the house bee (nurse bee) is accomplished in a regurgitated behavior. The role of the house bee is converting that nectar to honey or royal jelly (bee food) in a secreting process. In this synthesizing honey process, the main work is to split the complex sucrose sugar into fructose and glucose, which are simpler sugars and predominant in honey. This sucrose-splitting process is performed by adding the invertase, which is a special enzyme, to the nectar from the hypopharyngeal gland in the head of bee. Then, the new synthesized honey or royal jelly is spread out in a honey comb cells. The house bee exposes this secreted honey as a thin film to aware of the last filtration. This final step was done by increasing the surface area, to insure the fast evaporation of water in the well-done honey. Finally, the filled honey comb cells sealed and capped by propolis (plant gum), which is an adhesive material. This waxy cover prevents the honey from the bacterial attacks or in case of prevention the stored food to avoid the fermentation.

Consequently, here the details of the foraging process are presented to make a base for our nature-inspired method. It is a hybrid adaptation from the process of honeybees in collecting nectar to make honey and royal jelly. The proposed ABCRna method solves the secondary structure prediction problem of RNA with pseudoknots. The idea is stimulating a hybrid novelty swarm-intelligence approach from collecting nectar and making honey in the natural secretion process. ABCRna as a new optimization algorithm is based on the main features of a hybrid between two heuristic-based method KnotSeeker [3] and dynamic programming algorithms UNAFold [26].

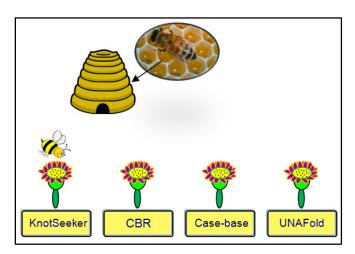


Figure 4. A secreting honey process: the simulation of honeybees collecting nectar to secret honey and royal jelly mapping with a diagrammatic representation of the proposed method.

E. CBR and KB

It's commonly known that the AI research area provides many methodologies and technologies for solving complex problems, which the CBR is one of them. Recently, the CBR has been successfully used to restore solution for a new problem based on expertise by retrieving the similar mature solutions of the past problems [55]. Originally, CBR comes up from the cognitive science and the human expertise to retain and retrieve the information. In another word in CBR method, the people solve the new problem by recalling how they solved the past similar problems. The CBR method includes a problem solving cycle with four main activities: Retrieve, Reuse, Revise and Retain [56]. According to the Figure 5, in the heart of this four-RE's cycle there is a caselibrary as a Knowledge Base (KB). This KB is used in retrieval action to assess an intelligent decision of the similar cases for revising the final outputs by retrieving the most correct solutions.

By referring to the adhere of exact matching concepts, the CBR is a generic AI methodology in problem solving [57]. In the proposed ABCRna method, the CBR is deployed as a modern AI inspired technique with KB to augment the result in retrieval steps. The role of CBR is finding the current pseudoknotted RNA sub-structure with the exact matching from KB. The KB holds and clusters all real pseudoknotted RNA sequences and their known native structures. If the retrieval one has pseudoknots in its secondary structure, then the CBR chooses the current one. This CBR comparing process, enhances the quality of the predicted pseudoknotted RNA secondary structure. Moreover, it is deployed significantly to be an alternative development technique for solving the secondary structure prediction problem of RNA with pseudoknots.

F. Preliminarly in Optimization

According to the theoretical viewpoint, the optimization methods are branch of the applied mathematics and basically

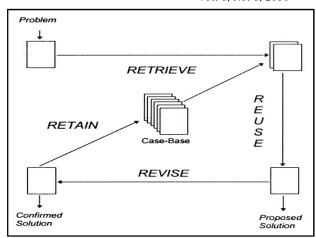


Figure 5. The Case-Based Reasoning (CBR), a modern Artificial Intelligence methodology, adapted from [55].

compromise from two main classes of algorithms; deterministic and probabilistic. Figure 6 shows the general category of the global optimization methods to clear the relation among all their characteristics. Definitely, the deterministic algorithms are a type of algorithm which take a set of fixed inputs and produce a fixed result. While, the heuristic is a single assumption works as a search strategy or technique in problem-solving. It is based on intelligence and experience, which can be applied loosely in computer implementation [58]. The meta-heuristic is based-on several assumptions work as an optimizer to improve a series of candidate solutions to reach to the final problem solving. Also it may use the many trials iteratively. In 2001, Geem et al. introduced the Harmony Search (HS) algorithm, which was a new meta-heuristic algorithm based on natural-inspired phenomena behavioral models [59]. The HS has been developed from mimicking the natural phenomena of the players). musicians improvisation (music Several experiments proved that the HS as a meta-heuristic algorithm, is capable to solve the optimization problems with more improved performance. The result makes the HS as a durable meta-heuristic algorithm in solving the NP-complete problems. The Traveling Salesman Problem (TSP) is an example of NP-problem which was solved by HS [60].

Now the main question, Is it feasible to develop a hybrid meta-heuristic algorithm for building the pseudoknotted RNA structure with good performance and more accurate result? To do this an optimized swarm-based intelligence algorithm would be inspired as a kind of stimulation from the Artificial Bee Colony (ABC) algorithm [61]. This inspired proposal utilizes the ABC to solve the related issue of RNA structure in bioinformatics. Moreover, the Particle Swarm Optimization (PSO) is a distinguished swarm-based intelligence algorithm that models some animal social behavior like fish schooling or swarm of honey-bees [62]. PSO has been proposed by Kennedy in 1995 and has reached to be an interesting area of knowledge to exploit for developing a new meta-heuristic algorithm by mimicking and inspiring the natural phenomena of animals and colony insects.

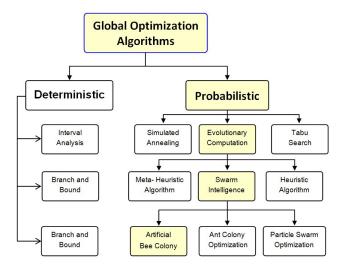


Figure 6. A schematic diagram of the global optimization methods.

IV. PROPOSED METHOD

This section explains in details, the new hybrid of derived ABC algorithm to overcome the pseudoknotted RNA secondary structure prediction problem. The proposed hybrid ABCRna method is inspired from the swarm-intelligence social behavioral model of honey-bees in collecting nectar and secreting honey, as shown in Figure 7. Hence, the authors develop ABCRna as a hybrid method in a simple way to build the secondary structure of RNA molecule with pseudoknots. The following sub-sections demonstrate separately the paradigms of designing the proposed method. These sub-sections describe the mapping of the all features between the ABC optimized algorithm and the RNA structural prediction problem. The final computational results of ABCRna for RNA structure reveal an optimized better performance and more accuracy in terms of sensitivity and specificity. Its computer code implementation shows less space and time complexities when comparing with other state-of-the-art methods in solving such RNA prediction problem.

Here, the researchers underline the hybrid adaption model as a new derivation from ABC algorithm to solve RNA prediction problem. It is a first threshold further opens the door in front of the other bioinformatics researchers to follow. Furthermore, it gives immense opportunity to expand this proposed optimizer in solving such kind of complex biocomputing problems. This is why the AI material already has presented in the background section to be a general guidance.

A. Honey-bee Foraging Algorithm

The innovative ABC as a swarm-based intelligence algorithm was deployed particularly based on the honeybee natural social behaviours. A few other algorithms have been derived by inspiring the honeybees swarm behavioral model, intelligently [61]. Many researchers have been adapted such this swarm collective behaviours to solve optimization combinatorial problems. Herein, we describe a new hybrid

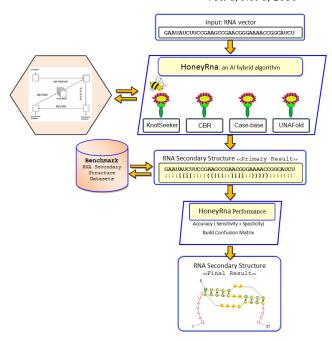


Figure 7. Workflow of the ABCRna approach for predicting the pseudoknotted RNA secondary structure, some parts adapted [55].

algorithm called "ABCRna", which is derived from the original ABC algorithm [24]. It is developed as a hybrid adaptation between ABC models with deterministic constraints and inspired by the intelligence social behaviours of bees in collecting nectar to secret honey. The proposed method is applied to solve the pseudoknotted RNA secondary structure prediction problem, which is a kind of combinatorial NP-complete problem [20].

The bees in colony deliberated for collecting nectar and secreting honey and they compromise in three bee groups: employed bees, unemployed bees (onlookers or scouts) and nurse bees, plus the food sources (flower patches profitability). The first two groups of honeybees (employed and unemployed) search for the last part which is the rich food sources. The third bee component takes the collected forage (nectar) from the first two groups by process of regurgitation. After that, the nurse bee starts making honey and royal jelly by a popular secretion honey process. The behavioral steps of the bees to carry out the forage collecting process, has been shown in Figure 4. Naturally, it can be described as follows:

- a) Employed bee (Forager): visits several food sources to collect the harvested crop, in each round-trip foraging journey. Nectar from many flower patches accumulate and store in the foragers' honey stomach (honey sac).
- b) Nurse bee: working inside the hive and she receive the collected nectar from the employed bee (forager) by regurgitation process. After that, the nurse bee starting makes honey or royal jelly from the associated mixed nectar by secreting invertase enzyme from the hypopharyngeal gland in her head. The corresponding enzyme assists to split the

complex sugar (sucrose) to two simplifier sugars (fructose and glucose), which are principal of new well-done honey.

B. The Classical ABC Algorithm

The ABC algorithm is a new AI model, which has been stimulated by the collective behavior of the social honeybees based on swam intelligence. It uses multi-resource and multi-form to perform the job with full optimization [24]. The ABC algorithm originally is divided into three parts: employed bees, onlookers and scouts. The employed bee is a hard-worker part in the colony that responsible to collect food. Onlookers part is waiting inside the hive to decide on a forage source. Scouts is performed a general search to find the food resources.

C. Hybrid ABC Algorithm for RNA Structural Prediction

Our proposed ABCRna method is a hybrid model based on the PSO and it is derived from the original ABC model. This new derivation of the modified ABC is associated with a specific case corresponding to the pseudoknotted RNA secondary structure prediction problem. The worker bee (employed bee) works as an agent, visits many rich flower patches (artificial food sources) to collect the nectar. Thereafter, all collected nectar from many flowers stored in foragers' honey stomach, which will be a mixture of nectar from several food sources (many flowers). Then, the employed bee (forager) back to the hive from the foraging journey with the mixed nectar fills her honey stomach. In the hive, the forager submits the crop (collected nectar) to the nurse bee in a regurgitation process. Finally, the nurse bee now is ready to make honey from the corresponding mixture of gathered nectar that submitted by employed bee. The nurse bee starts secreting the honey or royal jelly associated with specific needs of the hive. Here, the final well-done honey is represented the good solutions for the RNA structural prediction problem. In another words, the concluding honey in mapping segment, stands for the more accurate pseudoknotted RNA secondary structure for a giving primary sequence.

The central phase of the ABCRna method is a HoneyRna algorithm, which is a modified from ABC algorithm to solve the pseudoknotted RNA structural prediction method. This HoneyRna algorithm is illustrated in Figure 7 and computes in steps as follows:

1: Initialize

2: **REPEAT**

- 3: Place the employed bee on her food sources (many flowers)
- 4: Place the nurse bee on hive working to receive mixed nectar
- 5: Secret enzyme to split the complex nectar to a simpler honey
- 6: Fill the secreting food (honey & royal jelly) in the honeycomb
- 7: Filter the well-done honey from extra water by evaporation
- 8: Cap and seal the filled cell with food by adhesive wax
- 9: UNTIL (Demanded food is met)

In the modified ABC algorithm, each cycle of the collecting nectar and secreting honey process includes three

steps: (i) the employed bee visits many flower patches in each round-trip of collecting nectar journey. All gathered nectar is stored in her honey stomach. The employed bee backs to the hive with holding the mixed nectar. Then, she will submit this mixture to the nurse bee.

Moreover, the secretion process of the honey by nurse bee performs in many steps, as follows:

- a) The harvest of the forage (the mixed nectar), has been collected from many flowers. By mapping this phase with the RNA related issue, the predicted RNA secondary structure is collected from many existed RNA predicted methods, as illustrated in Figure 7.
- b) The nurse bee starts make honey by secreting invertase enzyme from the gland in her head. This enzyme simplifies the sucrose which is a complex sugar in the nectar to two types (fructose and glucose) of simpler sugars, which are composed the well done honey. By mapping this with RNA structural problem, there is an agent program, which is working like that enzyme. This function re-constructs the entire secondary structure of RNA sequence with pseudoknots from many parts.

V. BENCHMARK TESTS AND RESULTS

We evaluated ABCRna on different types of pseudoknotted RNA classes. The proposed method is built to predict the RNA secondary structure with pseudoknots. The comparisons of the ABCRna results have been performed by measuring the accuracy of its outputs to the outputs that has been achieved from FlexStem [29], HotKnots [30] and pknotsRG [19]. These accuracy measurements compromise three statistical notations: (Sensitivity *S*, Selectivity *P* and F-measure). They can be calculated by applying the following formulas, which derived from [63]:

Sensitivity
$$S = \frac{TP}{TP + FN} \times 100$$
 (1)

Specificity
$$P = \frac{TP}{TP + FP} \times 100$$
 (2)

$$F - measure = 2 \times P \times (\frac{S}{P + S}) \times 100, \quad (3)$$

where TP is represented the True Positive, which denotes the number of base pairs that are predicted correctly and presented in the known native structure. FN is represented the False Negative, which counts the base pairs that are presented in the known native structure, but they are not reported in the predicted structure. FP is represented the False positive, which denotes the number of base pairs, presented in the native known structure, but they are not in predicted structure. F-measure is a single measure that combines both sensitivity and specificity of the predictor algorithm in a unique performance measure.

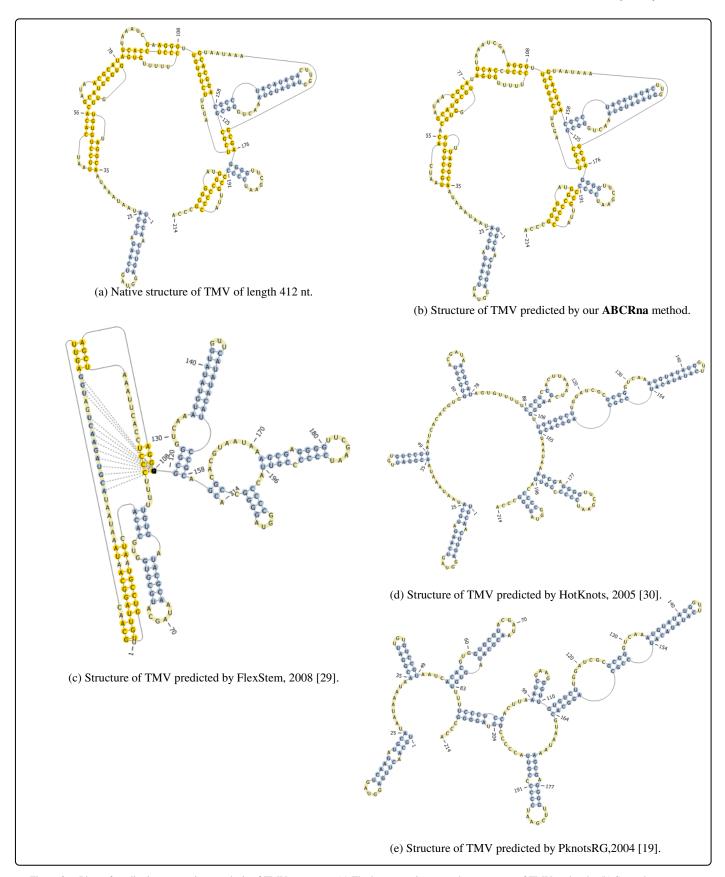


Figure 8. Plots of qualitative comparison analysis of TMV structures: (a) The known native secondary structure of TMV molecule. (b) Secondary structure predicted by our proposed ABCRna method, with highest excellent sensitivity of (92.9%) and specificity (95.6%). (c) Secondary structure predicted by FlexStem (sensitivity of 44.3% and specificity 44.9%). (d) Secondary structure predicted by HotKnots (sensitivity of 67.1% and specificity 81.0%). (e) Secondary structure predicted by pknotsRG (sensitivity of 60.0% and specificity 66.7%).

Here, the comparison analysis of the outputs are performed between our proposed ABCRna method against to the FlexStem [29], HotKnots [30] and pknotsRG [19]. One example of this comparison pricess uses the RNA sequence tobacco mosaic virus (TMV) from 3'UTR type [64]. The length of TMV is equal 214 nucleotides (nt), which its accession number "J02415". Our proposed ABCRna method obtained the highest results, Sensitivity (S = 92.9%) and Specificity (P = 95.6%), which are measured according to the known native structure of TMV molecule. The sensitivity and specificity of FlexStem [29], HotKnots [30] and pknotsRG [19], are listed in the legend of the illustrative Figure 8, respectively. Finally, the Figure 8 depicts a qualitative comparison analysis among the output of our ABCRna and the best result of all others methods from the literature. This comparison analysis is applied on the secondary structure of TMV RNA molecule, which its images are produced by PseudoViewer software tool [65]. NUPACK [66] and pknotsRE [67] methods cannot predict the secondary structure of RNA sequences in larger than the length of 200 nt and 150 nt, respectively. The reason behind that the both algorithms NUPACK [66] and pknotsRE [67], require an enormous amount of memory (RAM) and run in exponential time. To reach to a fair comparison, the outputs for these scenarios put out of the result. Also, all five existing methods (FlexStem [29], HotKnots [30], pknotsRG [19], NUPACK [66] and pknotsRE [67]), have been implemented in the same machine, a PC Ubuntu 10.04 64-bit Linux OS, with AMD Phenom-II 810 2.6-GHz Quad-Core processor and Dual Channel 4GB (2x2GB) DDR2-800 Memory (RAM).

Table 1 summarizes the final comparison analysis of the results among the predicted RNA structures from our proposed ABCRna method and the best ones from FlexStem [29], HotKnots [30], pknotsRG [19], NUPACK [66] and pknotsRE [67] methods. The comparison process has been done in respects to the three accuracy metrics listed in Equations (1, 2 and 3). The evaluation of these comparative results were performed and verified according to the standard native structures of each RNA molecule. The analyses show the results of the ABCRna method are significantly better than the results of other methods from literature, in terms of sensitivity, specificity and F-measure.

TABLE I. COMPARISON ACCURACY METRICS THE STRUCTURAL RESULTS OF BASE PAIRS AMONG THE ABCRNA METHOD AND OTHER METHODS.

RNA sequence				ABC	CRna (2	2010)	FlexStem (2008)			HotKnots (2005)		pknotsRG (2004)		NUPACK (2003)			pknotsRE (1999)				
Seq. ID, Ref.	RNA class	Accession Number	Length (nt)	S	P	F	s	P	F	s	P	F	s	P	F	s	P	F	s	P	F
HDV-It_g [68]	Ribozymes	X04451	87	90.6	93.5	92.1	84.4	79.4	81.8	71.9	82.1	76.7	90.6	93.5	92.1	62.5	62.5	62.5	81.3	81.3	81.3
BMV [69]	3'-UTR	J02042	145	73.8	62.0	67.4	45.2	38.8	41.8	31.0	31.7	31.3	47.6	43.5	45.5	45.2	44.2	44.7	45.2	44.2	44.7
FMDV-A [70]	5'-UTR	AY593751	165	83.3	54.1	65.6	66.7	41.0	50.8	12.5	07.3	09.2	66.7	41.0	50.8	62.5	34.9	44.8	*	*	*
TMV [71]	3'-UTR	J02415	214	92.9	95.6	94.2	44.3	44.9	44.6	67.1	81.0	73.4	60.0	66.7	63.2	*	*	*	*	*	*

The highest results are in **bold**, and "*" denotes that the algorithm is unable to complete the run.

IV. CONCLUSION AND FUTURE WORK

This paper presented a novel hybrid method for solving RNA secondary structure with pseudoknot functional classes. This hybrid method includes ABC model as a global optimization method, hybridized with CBR as a local optimization technique. The proposed method used the existing results from KnotSeeker and UNAFold to generate a secondary structure of RNA includes pseudoknots, by using an existing cases.

Three evaluation mechanisms are used to measure the efficiency and performance of proposed method comparing to others from literature. The sensitivity, specificity and F-measure metrics showed that successful outcomes have been recorded. Furthermore, three different comparative methods are used in order to compare the obtained results. The proposed ABCRna hybrid method outperformed other comparators in almost all standard benchmarks. Note that the factor for comparison is the real native structure.

We believe that the proposed hybrid method have a high potential with a great efficiency for the problems solved by RNA community. This worthwhile domain is pregnant with several future research directions such as: further study cases in CBR, different global optimization, different factors of analysis and hybridize more RNA prominent methods.

ACKNOWLEDGMENT

This research was partly supported by a Universiti Sains Malaysia (USM) Fellowship, awarded to the corresponding author. It was also funded by "Insentif APEX". The authors gratefully acknowledge the reviewers for their helpful and useful comments.

REFERENCES

[1] R. Williams and O. Luo, "Complexity, Post-genomic Biology and Gene Expression Programs," *Complex Physical, Biophysical and Econophysical Systems: Proceedings of the 22nd Canberra International Physics Summer School*, p. 319, 2009.

- [2] F. Buchholz, R. Kittler, M. Slabicki, and M. Theis, "Enzymatically prepared RNAi libraries," *Nat Meth*, vol. 3, pp. 696-700, 2006.
- [3] J. Sperschneider and A. Datta, "KnotSeeker: Heuristic pseudoknot detection in long RNA sequences," RNA, vol. 14, pp. 630-640, April 2008 2008.
- [4] G. Sengle, A. Eisenführ, P. S. Arora, J. S. Nowick, and M. Famulok, "Novel RNA catalysts for the Michael reaction," *Chemistry & Biology*, vol. 8, pp. 459-473, 2001.
- [5] J. Mattick, "RNA regulation: a new genetics?," *Nature Reviews Genetics*, vol. 5, pp. 316-323, 2004.
- [6] O. Nureki, D. Vassylyev, M. Tateno, A. Shimada, T. Nakama, S. Fukai, M. Konno, T. Hendrickson, P. Schimmel, and S. Yokoyama, "Enzyme structure with two catalytic sites for double-sieve selection of substrate," *Science*, vol. 280, p. 578, 1998.
- [7] R. Rambo and J. Tainer, "Bridging the solution divide: comprehensive structural analyses of dynamic RNA, DNA, and protein assemblies by small-angle X-ray scattering," *Current Opinion in Structural Biology*, vol. 20, pp. 128-137, 2010.
- [8] A. Heck, "Native mass spectrometry: a bridge between interactomics and structural biology," *Nature Methods*, vol. 5, pp. 927-933, 2008.
- [9] H.-K. Cheong, E. Hwang, C. Lee, B.-S. Choi, and C. Cheong, "Rapid preparation of RNA samples for NMR spectroscopy and X-ray crystallography," *Nucl. Acids Res.*, vol. 32, pp. e84-, June 15, 2004 2004.
- [10] R. M. Al-Khatib, R. Abdullah, and N. Abdul Rashid, "A Comparative Taxonomy of Parallel Algorithms for RNA Secondary Structure Prediction," *Evolutionary Bioinformatics*, vol. 6, p. 27, 2010.
- [11] C. Wilson, J. Nix, and J. Szostak, "Functional Requirements for Specific Ligand Recognition by a Biotin-Binding RNA Pseudoknot†," *Biochemistry*, vol. 37, pp. 14410-14419, 1998.
- [12] A. Roth and R. Breaker, "The structural and functional diversity of metabolite-binding riboswitches," *Annual review of biochemistry*, vol. 78, pp. 305-334, 2009.
- [13] Y. Takakura, "Towards therapeutic application of RNA-mediated gene regulation," *Advanced Drug Delivery Reviews*, vol. 61, pp. 667-667, 2000
- [14] P. Menzel, J. Gorodkin, and P. F. Stadler, "The tedious task of finding homologous noncoding RNA genes," RNA, vol. 15, pp. 2075-2082, December 2009 2009.
- [15] R. M. Al-Khatib, R. Abdullah, and N. Abdul Rashid, "A Survey of Compute Intensive Algorithms for Ribo Nucleic Acids Structural Detection," *Journal of Computer Science*, vol. 5, pp. 680-689, 2009.
- [16] C. W. A. Pleij, K. Rietveld, and L. Bosch, "A new principle of RNA folding based on pseudoknotting," *Nucl. Acids Res.*, vol. 13, pp. 1717-1731, March 11, 1985 1985.
- [17] M. Zuker, "Mfold web server for nucleic acid folding and hybridization prediction," *Nucl. Acids Res.*, vol. 31, pp. 3406-3415, July 1, 2003 2003.
- [18] I. L. Hofacker, "Vienna RNA secondary structure server," Nucl. Acids Res., vol. 31, pp. 3429-3431, July 1, 2003 2003.
- [19] J. Reeder and R. Giegerich, "Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics," BMC Bioinformatics, vol. 5, p. 104, 2004.
- [20] T. Akutsu, "Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots," *Discrete Applied Mathematics*, vol. 104, pp. 45-62, 2000.
- [21] R. B. Lyngso and C. N. S. Pedersen, "RNA Pseudoknot Prediction in Energy-Based Models," *Journal of Computational Biology*, vol. 7, pp. 409-427, 2000.
- [22] Q. Dong and Z. Wu, "A linear-time algorithm for solving the molecular distance geometry problem with exact inter-atomic distances," *Journal of Global Optimization*, vol. 22, pp. 365-375, 2002.
 [23] U. Schning and M. von Knop, "Using Stochastic Indexed Grammars for Distances of the Contraction of the Con
- [23] U. Schning and M. von Knop, "Using Stochastic Indexed Grammars for RNA Structure PredictionWith Pseudoknots," *Bulletin of the EATCS*, p. 185, 2010.
- [24] D. Karaboga, "An idea based on honey bee swarm for numerical optimization," in *Technical Report-TR06, Erciyes University, Engineering Faculty, Computer Engineering Department*, 2005.
- [25] H. Chang, "Converging Marriage in Honey-Bees Optimization and Application to Stochastic Dynamic Programming," *Journal of Global Optimization*, vol. 35, pp. 423-441, 2006.
- [26] N. Markham and M. Zuker, "DINAMelt web server for nucleic acid melting prediction," *Nucleic acids research*, vol. 33, p. W577, 2005.
- [27] T. Furusawa, R. Rakwal, H. W. Nam, J. Shibato, G. K. Agrawal, Y. S. Kim, Y. Ogawa, Y. Yoshida, Y. Kouzuma, Y. Masuo, and M. Yonekura, "Comprehensive Royal Jelly (RJ) Proteomics Using One- and Two-Dimensional Proteomics Platforms Reveals Novel RJ Proteins and

- Potential Phospho/Glycoproteins," *Journal of Proteome Research*, vol. 7, pp. 3194-3229, 2008.
- [28] A. G. O. Yeh and X. Shi, "Case-based reasoning (CBR) in development control," *International Journal of Applied Earth Observation and Geoinformation*, vol. 3, pp. 238-251, 2001.
- [29] X. Chen, S. He, D. Bu, F. Zhang, Z. Wang, R. Chen, and W. Gao, "FlexStem: improving predictions of RNA secondary structures with pseudoknots by reducing the search space," *Bioinformatics*, vol. 24, p. 1994, 2008.
- [30] J. REN, B. RASTEGARI, A. CONDON, and H. H. HOOS, "HotKnots: Heuristic prediction of RNA secondary structures including pseudoknots," RNA, vol. 11, pp. 1494-1504, October 2005 2005.
- [31] W. Tichelaar, W. G. SCHUTTER, A. C. ARNBERG, E. F. J. BRUGGEN, and W. STENDER, "The quaternary structure of *Escherichia coli* RNA polymerase studied with (scanning) transmission (immuno)electron microscopy," *European Journal of Biochemistry*, vol. 135, pp. 263-269, 1983.
- [32] H. Liu, D. Xu, J. Shao, and Y. Wang, "An RNA folding algorithm including pseudoknots based on dynamic weighted matching," *Computational Biology and Chemistry*, vol. 30, pp. 72-76, 2006.
- [33] E. Westhof, "Toward atomic accuracy in RNA design," Nat Meth, vol. 7, pp. 272-273, 2010.
- [34] T. Xia, J. SantaLucia, M. E. Burkard, R. Kierzek, S. J. Schroeder, X. Jiao, C. Cox, and D. H. Turner, "Thermodynamic Parameters for an Expanded Nearest-Neighbor Model for Formation of RNA Duplexes with Watson-Crick Base Pairs," *Biochemistry*, vol. 37, pp. 14719-14735, 1998.
- [35] O. Schrader, T. Baumstark, and D. Riesner, "A Mini-RNA containing the tetraloop, wobble-pair and loop E motifs of the central conserved region of potato spindle tuber viroid is processed into a minicircle," *Nucl. Acids Res.*, vol. 31, pp. 988-998, February 1, 2003 2003.
- [36] A. Bar-Shira, A. Panet, and A. Honigman, "An RNA secondary structure juxtaposes two remote genetic signals for human T-cell leukemia virus type I RNA 3'-end processing," *J. Virol.*, vol. 65, pp. 5165-5173, October 1, 1991 1991.
- [37] I. Brierley, S. Pennell, and R. J. C. Gilbert, "Viral RNA pseudoknots: versatile motifs in gene expression and replication," *Nat Rev Micro*, vol. 5, pp. 598-610, 2007.
- [38] M. N. Win and C. D. Smolke, "A modular and extensible RNA-based gene-regulatory platform for engineering cellular function," *Proceedings* of the National Academy of Sciences, vol. 104, pp. 14283-14288, September 4, 2007 2007.
- [39] A. Mathuriya, D. A. Bader, C. E. Heitsch, and S. C. Harvey, "GTfold: a scalable multicore code for RNA secondary structure prediction," in Proceedings of the 2009 ACM symposium on Applied Computing Honolulu, Hawaii: ACM, 2009.
- [40] M. Masiero, G. Nardo, S. Indraccolo, and E. Favaro, "RNA interference: Implications for cancer treatment," *Molecular Aspects of Medicine*, vol. 28, pp. 143-166, 2007.
- [41] T. Schmickl, H. Hamann, H. Wrn, and K. Crailsheim, "Two different approaches to a macroscopic model of a bio-inspired robotic swarm," *Robotics and Autonomous Systems*, vol. 57, pp. 913-921, 2009.
- [42] D. Hill and T. Webster, "Apiculture and forestry (bees and trees)," Agroforestry Systems, vol. 29, pp. 313-320, 1995.
- [43] S. Ziane and A. Melouk, "A swarm intelligent multi-path routing for multimedia traffic over mobile ad hoc networks," in *Proceedings of the* 1st ACM international workshop on Quality of service & amp; security in wireless and mobile networks Montreal, Quebec, Canada: ACM, 2005.
- [44] Z. Lian, B. Jiao, and X. Gu, "A similar particle swarm optimization algorithm for job-shop scheduling to minimize makespan," *Applied Mathematics and Computation*, vol. 183, pp. 1008-1017, 2006.
- [45] W. BRYAN, S. ADRIAN, R. DIRK, and O. JAMES, "UAV Swarm Control: Calculating Digital Pheromone Fields with the GPU," *The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology*, vol. 3, pp. 167-176, 2006.
- [46] P. Doherty, P. Haslum, F. Heintz, T. Merz, P. Nyblom, T. Persson, and B. Wingman, "A Distributed Architecture for Autonomous Unmanned Aerial Vehicle Experimentation," in *Distributed Autonomous Robotic* Systems 6, 2007, pp. 233-242.
- [47] E. Bonabeau, M. Dorigo, and G. Theraulaz, "Inspiration for optimization from social insect behaviour," *Nature*, vol. 406, pp. 39-42, 2000.
- [48] S. Remolina and K. Hughes, "Evolution and mechanisms of long life and high fertility in queen honey bees," AGE, vol. 30, pp. 177-185, 2008.
- [49] D. Karaboga and B. Akay, "A comparative study of Artificial Bee Colony algorithm," *Applied Mathematics and Computation*, vol. 214, pp. 108-132, 2009.

Vol. 8, No. 8, 2010

- [50] I. Cakmak, C. Sanderson, T. D. Blocker, L. Lisa Pham, S. Checotah, A. A. Norman, B. K. Harader-Pate, R. Tyler Reidenbaugh, P. Nenchev, J. F. Barthell, and H. Wells, "Different solutions by bees to a foraging problem," *Animal Behaviour*, vol. 77, pp. 1273-1280, 2009.
- [51] T. D. Seeley, "Social foraging by honeybees: how colonies allocate foragers among patches of flowers," *Behavioral Ecology and Sociobiology*, vol. 19, pp. 343-354, 1986.
- [52] T. D. Seeley, A. S. Mikheyev, and G. J. Pagano, "Dancing bees tune both duration and rate of waggle-run production in relation to nectar-source profitability," *Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology,* vol. 186, pp. 813-819, 2000.
- [53] K. R. Abbott and R. Dukas, "Honeybees consider flower danger in their waggle dance," *Animal Behaviour*, vol. 78, pp. 633-635, 2009.
- [54] T. Kubo, M. Sasaki, J. Nakamura, H. Sasagawa, K. Ohashi, H. Takeuchi, and S. Natori, "Change in the Expression of Hypopharyngeal-Gland Proteins of the Worker Honeybees (Apis melliferaL.) with Age and/or Role," *J Biochem*, vol. 119, pp. 291-295, February 1, 1996 1996.
- [55] A. Aamodt and E. Plaza, "Case-based reasoning: foundational issues, methodological variations, and system approaches," *AI Commun.*, vol. 7, pp. 39-59, 1994.
 [56] I. Watson, "Case-based reasoning is a methodology not a technology,"
- [56] I. Watson, "Case-based reasoning is a methodology not a technology," Knowledge-Based Systems, vol. 12, pp. 303-308, 1999.
- [57] L. Portinale, D. Magro, and P. Torasso, "Multi-modal diagnosis combining case-based and model-based reasoning: a formal and experimental analysis," *Artificial Intelligence*, vol. 158, pp. 109-153, 2004.
- [58] J. Pearl, Heuristics: Intelligent search strategies for computer problem solving, 1984.
- [59] Z. Geem, J. Kim, and G. V. Loganathan, "A New Heuristic Optimization Algorithm: Harmony Search," SIMULATION, vol. 76, pp. 60-68, 2001.
- [60] K. S. Lee and Z. W. Geem, "A new structural optimization method based on the harmony search algorithm," *Computers & Structures*, vol. 82, pp. 781-798, 2004
- [61] D. Karaboga and B. Basturk, "On the performance of artificial bee colony (ABC) algorithm," Applied Soft Computing, vol. 8, pp. 687-697, 2008.



Ra'ed M. Al-Khatib received his Bachelors degree in Computer Science from Mu'tah University, Karak, Jordan in 1993 and Masters Degree in Computer Engineering-Embedded System from Yarmouk University, Irbid, Jordan in 2006. He is currently a researcher at the Parallel and Distributed Processing Research Group and a PhD candidate as well under the supervision of Professor Dr. Rosni Abdullah and Associated Professor Dr. Nur'Aini Abdul Rashid at the School of Computer Sciences, Universiti Sains Malaysia in the area of Parallel Algorithms Applied to Bioinformatics Applications.



Rosni Abdullah received her Bachelor's Degree in Computer Science and Applied Mathematics and Masters Degree in Computer Science from Western Michigan University, Kalamazoo, Michigan, U.S.A. in 1984 and 1986 respectively. She joined the School of Computer Sciences at Universiti Sains Malaysia in 1987 as a lecturer. She received an award from USM in 1993 to pursue her PhD at Loughborough University in United Kingdom in the area Parallel Algorithms. She was promoted to Associate Professor in 2000 and to Professor in 2008. She has held several administrative positions such as First Year Coordinator, Programme Chairman and Deputy Dean for Postgraduate Studies and Research. She is currently the Dean of the School of Computer Sciences and also Head of the Parallel and Distributed Processing Research Group which focus on grid computing and bioinformatics research. Her current research work is in the area of Parallel Algorithms for Bioinformatics Applications.

- [62] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Neural Networks*, 1995. Proceedings., IEEE International Conference on, 1995, pp. 1942-1948 vol.4.
- [63] P. Baldi, S. Brunak, Y. Chauvin, C. Andersen, and H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: an overview," *Bioinformatics*, vol. 16, p. 412, 2000.
- [64] F. H. D. van Batenburg, A. P. Gultyaev, and C. W. A. Pleij, "PseudoBase: structural information on RNA pseudoknots," *Nucl. Acids Res.*, vol. 29, pp. 194-195, January 1, 2001 2001.
- [65] Y. Byun and K. Han, "PseudoViewer3: generating planar drawings of large-scale RNA structures with pseudoknots," *Bioinformatics*, vol. 25, p. 1435, 2009.
- [66] R. Dirks and N. Pierce, "A partition function algorithm for nucleic acid secondary structure including pseudoknots," *J Comput Chem*, vol. 24, pp. 1664 - 1677, 2003.
- [67] E. Rivas and S. R. Eddy, "A dynamic programming algorithm for RNA structure prediction including pseudoknots," *Journal of Molecular Biology*, vol. 285, pp. 2053-2068, 1999.
- [68] M. D. Been and G. S. Wickham, "Self-Cleaving Ribozymes of Hepatitis Delta Virus RNA," *European Journal of Biochemistry*, vol. 247, pp. 741-753, 1997.
- [69] C. W. A. Pleij, J. P. Abrahams, A. v. Belkum, K. Rietveld, and L. Bosch, "The spatial folding of the 3' noncoding region of aminoacylatable plant viral RNAs," M. Brinton and R. Ruecker, Editors, Positive Strand RNA Viruses, A. R. Liss, New York (1987), pp. 299–316, 1987 1987.
- [70] B. E. Clarke, A. L. Brown, K. M. Currey, S. E. Newton, D. J. Rowlands, and A. R. Carroll, "Potential secondary and tertiary structure in the genomic RNA of foot and mouth disease virus," *Nucleic Acids Research*, vol. 15, pp. 7067-7079, September 11, 1987 1987.
- [71] R. Koenig, S. Barends, A. P. Gultyaev, D.-E. Lesemann, H. J. Vetten, S. Loss, and C. W. A. Pleij, "Nemesia ring necrosis virus: a new tymovirus with a genomic RNA having a histidylatable tobamovirus-like 3' end," *J Gen Virol*, vol. 86, pp. 1827-1833, June 1, 2005 2005.



Nur'Aini Abdul Rashid received her Bachelor's Degree in Computer Science from Mississippi State University, U.S.A. in 1985. She joined the School of Computer Sciences at Universiti Sains Malaysia in 1988 as a tutor. She received her Master's Degree in Computer Sciences from Universiti Sains Malaysia, Penang, Malaysia in 1995. She continued in the School of Computer Sciences at Universiti Sains Malaysia as a lecturer from 1995. She received an award from USM in 2002 to pursue her PhD at Universiti Sains Malaysia, Penang in the area of Parallel methods. She was promoted to Senior Lecturer in 2004. She is currently an Associate Professor in the School of Computer Sciences and also member of the Parallel and Distributed Processing Research Group which focus on grid computing and bioinformatics research. Her current research work is in the area of Parallel Algorithms applied to Bioinformatics Applications, Genomic Information Retrieval, Text Search and Comparison Methods.

Hybrid JPEG Compression Using Histogram Based Segmentation

M.Mohamed Sathik¹, K.Senthamarai Kannan² and Y.Jacob Vetha Raj³

¹Department of Computer Science, Sadakathullah Appa College, Tirunelveli, India.

Abstract-- Image compression is an inevitable solution for image transmission since the channel bandwidth is limited and the demand is for faster transmission. Storage limitation is also forcing to go for image compression as the color resolution and spatial resolutions are increasing according to quality requirements. JPEG compression is a widely used compression technique. JPEG compression method uses linear quantization and threshold values to maintain certain quality in an entire image. The proposed method estimates the vitality of the block of the image and adapts variable quantization and threshold values. This ensures that the vital area of the image is highly preserved than the other areas of the image. This hybrid approach increases the compression ratio and produces a desired high quality output image.

Key words-- Image Compression, Edge-Detection, Segmentation. Image Transformation, JPEG, Quantization.

I. INTRODUCTION

Every day, an enormous amount of information is stored, processed, and transmitted digitally. Companies provide business associates, investors, and potential customers with financial data, annual reports, inventory, and product information over the Internet. And much of the online information is graphical or pictorial in nature; the storage and communications requirements are immense. Methods of compressing the data prior to storage and/or transmission are of significant practical and commercial interest.

Compression techniques fall under two broad categories such as lossless[1] and lossy[2][3]. The former is particularly useful in image archiving and allows the image to be compressed and decompressed without losing any information. And the later, provides higher levels of data reduction but result in a less than perfect reproduction of the original image. Lossy compression is useful in applications such as broadcast television, videoconferencing, and facsimile transmission, in which certain amount of error is an acceptable trade-off for increased compression performance. The foremost aim of image compression is to reduce the number of bits needed to represent an image. In lossless image compression algorithms, the reconstructed image is identical to the original image. Lossless algorithms, however, are limited by the low compression ratios they can achieve. Lossy compression algorithms, on the other hand, are capable of achieving high compression ratios. Though the reconstructed image is not identical to the original image,

lossy compression algorithms obtain high compression ratios by exploiting human visual properties.

Vector quantization [4],[5], wavelet transformation [1], [5]-[10] techniques are widely used in addition to various other methods[11]-[17] in image compression. The problem in lossless compression is that, the compression ratio is very less; where as in the lossy compression the compression ratio is very high but may loose vital information of the image. Some of the works carried out in hybrid image compression [18]-[19] incorporated different compression schemes like PVQ and DCTVQ in a single image compression. But the proposed method uses lossy compression method with different quality levels based on the context to compress a single image by avoiding the difficulties of using side information for image decompression in [20].

The proposed method performs a hybrid compression, which makes a balance on compression ratio and image quality by compressing the vital parts of the image with high quality. In this approach the main subject in the image is very important than the background image. Considering the importance of image components, and the effect of smoothness in image compression, this method segments the image as main subject and background, then the background of the image is subjected to low quality lossy compression and the main subject is compressed with high quality lossy compression. There are enormous amount of work on image compression is carried out both in lossless [1] [14] [17] and lossy [4] [15] compression. Very few works are carried out for Hybrid Image compression [18]-[20].

In the proposed work, for image compression, the edge detection, segmentation, smoothing and dilation techniques are used. For edge detection, segmentation [21],[22] smoothing and dilation, there are lots of work has been carried out [2],[3]. A novel and a time efficient method to detect edges and segmentation used in the proposed work are described in section II, section III gives a detailed description of the proposed method, the results and discussion are given in section IV and the concluding remarks are given in section V.

II. BACKGROUND

A. JPEG Compression

Components of Image Compression System (JPEG). Image compression system consists of three closely connected components namely

• Source encoder (DCT based)

^{2,3}Department of Statistics, Manonmanium Sundaranar University, Tirunelveli, India.

- Quantizer
- Entropy encoder

Figure 2 shows the architecture of the JPEG encoder.

Principles behind JPEG Compression. A common characteristic of most images is that the neighboring pixels are correlated and therefore contain redundant information. The foremost task then is to find less correlated representation of the image. Two fundamental components of compression are redundancy and irrelevancy reduction. Redundancy reduction aims at removing duplication from the signal source. Irrelevancy reduction omits parts of the signal that will not be noticed by the signal receiver, namely the Human Visual System (HVS). The JPEG compression standard (DCT based) employs the use of the discrete cosine transform, which is applied to each 8 x 8 block of the partitioned image. Compression is then achieved by performing quantization of each of those 8 x 8 coefficient blocks.

$\begin{array}{lll} \textbf{Image} & \textbf{Transform} & \textbf{Coding} & \textbf{For} & \textbf{JPEG} & \textbf{Compression} \\ \textbf{Algorithm.} \end{array}$

In the image compression algorithm, the input image is divided into 8-by-8 or 16-by-16 non-overlapping blocks, and the two-dimensional DCT is computed for each block. The DCT coefficients are then quantized, coded, and transmitted. The JPEG receiver (or JPEG file reader) decodes the quantized DCT coefficients, computes the inverse two-dimensional DCT of each block, and then puts the blocks back together into a single image. For typical images, many of the DCT coefficients have values close to zero; these coefficients can be discarded without seriously affecting the quality of the reconstructed image. A two dimensional DCT of an M by N matrix A is defined as follows

$$B_{pq} = \alpha_p \alpha_q \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} A_{mn} \cos \frac{\pi (2m+1)p}{2M} \cos \frac{\pi (2n+1)q}{2N} , 0 \le p \le M-1$$

$$0 \le q \le N-1$$

where

$$\alpha_p = \begin{cases} 1/\sqrt{M} & p = 0\\ \sqrt{2/M} & 1 \le p \le M-1 \end{cases}$$

$$\mathbf{Q}q = \begin{cases} 1/\sqrt{N} & \text{, } q = 0\\ \sqrt{2/N} & \text{, } 1 \le q \le N-1 \end{cases}$$

The DCT is an invertible transformation and its inverse is given by

$$A_{\mathit{mn}} = \alpha_p \alpha_q \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \ B_{pq} \cos \frac{\pi (2m+1) p}{2M} \cos \frac{\pi (2n+1) q}{2N} \ , \ 0 \leq p \leq M-1$$

Where

$$\alpha_p = \begin{cases} 1/\sqrt{M}, p = 0\\ \sqrt{2/M}, 1 \le p \le M-1 \end{cases}$$

$$\Omega q = \begin{cases} 1/\sqrt{N} & , q = 0\\ \sqrt{2/N} & , 1 \le q \le N-1 \end{cases}$$

The DCT based encoder can be thought of as essentially compression of a stream of 8 X 8 blocks of image samples. Each 8 X 8 block makes its way through each processing step, and yields output in compressed form into the data stream. Because adjacent image pixels are highly correlated, the 'forward' DCT (FDCT) processing step lays the foundation for achieving data compression by concentrating most of the signal in the lower spatial frequencies. For a typical 8 X 8 sample block from a typical source image, most of the spatial frequencies have zero or near-zero amplitude and need not be encoded. In principle, the DCT introduces no loss to the source image samples; it merely transforms them to a domain in which they can be more efficiently encoded.

After output from the FDCT, each of the 64 DCT coefficients is uniformly quantized in conjunction with a carefully designed 64 – element Quantization Table. At the decoder, the quantized values are multiplied by the corresponding QT elements to recover the original unquantized values. After quantization, all of the quantized coefficients are ordered into the "zig-zag" sequence as shown in figure 1. This ordering helps to facilitate entropy encoding by placing low-frequency non-zero coefficients before high-frequency coefficients. The DC coefficient, which contains a significant fraction of the total image energy, is differentially encoded.

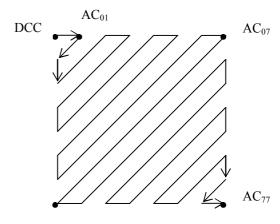


Figure 1 Zig-Zag Sequence

The JPEG decoder architecture is shown if figure 3 which is the reverse procedure described for compression.

B. Segmentation

Let D be a matrix of order m x n, represents the image of width m and height n. The domain for Di,j is [0..255], for any i=1..m, and any j=1..n.

The architecture of segmentation using histogram is shown in figure 4. To make the process faster the high resolution input image is down sampled 2 times. When the image is down sampled each time the dimension is reduced by half of the original dimension. So the final down sampled

image (D) is of the dimension $\frac{m}{4} \times \frac{n}{4}$. The down sampled image is smoothed to get smoothed gray scale image using equation (1).

equation (1).
$$S_{i,j} = 1/9(D_{i-1,,j-1} + D_{i-1,,j} + D_{i-1,,j+1} + D_{i,,j-1} + D_{i,,j} + D_{i,,j+1} + D_{i+1,,j-1} + D_{i+1,,j} + D_{i+1,,j+1})$$
 ...(1)

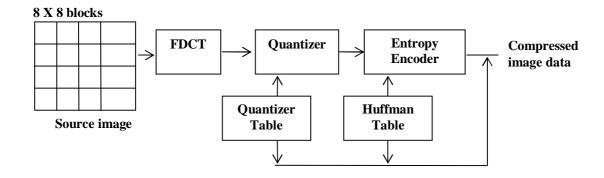


Figure 2. JPEG Encoder Block Diagram

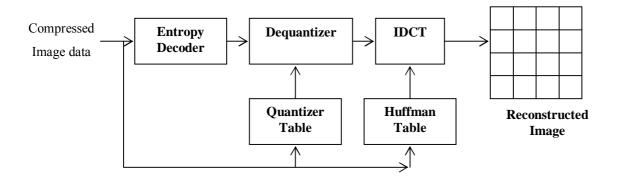


Figure 3. JPEG Decoder Block Diagram

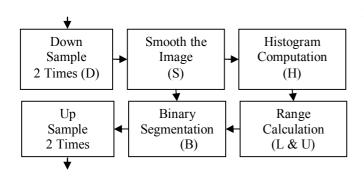


Figure -4 Classifier

The histogram(H) is computed for the gray scale image(S). The most frequently present gray scale value (Mh) is determined from the histogram by equation (2) and is shown as indicated by a line in figure 5.

$$Mh = arg\{max(H(x))\} \qquad (2)$$

The background value of the images is having the highest frequency in the case of homogenous background. In order to surmise background textures a range of gray level values are considered for segmentation. The range is computed using the equations (3) and (4).

$$L = max(Mh - 30,0)$$
 ...(3)

$$U = min(Mh + 30,255)$$
 ...(4)

The gray scale image S is segmented to detect the background area of the image using the function given in equation (5)

$$B_{i,j} = (S_{i,j} > L) \text{ and } (S_{i,j} < U)$$
 ...(5)

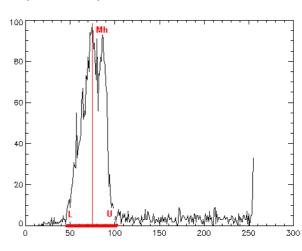


Figure – 5

After processing the pixel values for background area is 1 in the binary image B. To avoid the problem of over segmentation the binary image is subjected to sequence of morphological operations. The binary image is eroded with smaller circular structural element (SE) to remove smaller segments as given in equation (6).

$$B = B\Theta SE \qquad \dots (6)$$

Then the resultant image is subjected to morphological closing operation with larger circular structural element as given in equation (7).

$$B = B \bullet SE \qquad \dots (7)$$

III. PROPOSED HYBRID JPEG COMPRESSION.

The input image is initially segmented into background and foreground parts as described in section II.B Then the image is divided into 8x8 blocks and DCT values are computed for each block. The quantization is performed according to the predefined quantization table. The quantized values are then reordered based on zig-zag ordering method described in section II A. The lower values of AC coefficients are discarded from the zig-zag ordered list by comparing the threshold value selected by the selector as per the block's presences identified by the classifier. If the block is present in foreground area then the threshold is set to a higher value by the selector, otherwise a

- 10. Quantize the DCT coefficients
- 11. Discard lower quantized values based on the threshold value selected by the selector.
- 12. Compress remaining DCT coefficients by Entropy Encoder

The architecture of the proposed method is shown in figure 6. The Quantization Table is a fixed classical table derived from empirical results. The Quantizer quantizes the DCT coefficients computed by FDCT. The classifier identifies the class of each pixel by segmenting the given input image. The selector and limiter works together to find the discard threshold limit. The entropy encoder creates compressed code using the Code Table. The compressed image may be stored or transmitted faster than the existing method.

IV. RESULTS AND DISCUSSION

The Hybrid JPEG Compression method is implemented

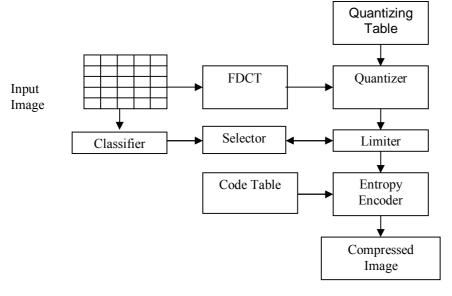


Figure 6. Hybrid JPEG compression method

lower value for threshold is set by the selector. After discarding insignificant coefficients the remaining data are compressed by the standard entropy encoder based on the code table.

Algorithm

- 1. Input High Resolution Color image.
- 2. Down sample the input image 2 times.
- 3. Convert the down sampled image to gray scale image (G).
- 4. Find histogram (H) of the gray scale image.
- 5. Find the lower (L) and upper (U) gray scale value of background area.
- 6. Find Binary segmented image (B) from the gray scale image (G)
- 7. Up sample Binary image (B) two times.
- 8. Divide the input image into 8x8 blocks
- 9. Find DCT coefficients for each blocks

according to the description in section III and tested with a set of test images shown in figure 8. The results obtained from the implementation of the proposed algorithms are shown in figures 7, 9,10 and table I. Figure 7.a shows the original input image. In Figure 7.b the segmented object and background area is discriminated by black and white. The compressed bit rates of the twelve test images are computed and tabulated in table 1. The low quality (LQ) and high quality (HQ) JPEG compression is performed and the corresponding compression ratios(CR) and PSNR values are tabulated. The PSNR is higher for HQ and CR is higher for LO. The Hybrid JPEG compression performs HO compression on main subject area and LQ compression on background area thus the PSNR value at main subject area is the same for Hybrid JPEG and HQ JPEG. Figure 9 shows the comparison of normalized CRs of Hybrid JPEG and HQ JPEG, it is observed that almost all of the images are compressed better than classical JPEG compression. Figure 10 shows how well the compression ratio is increased than the classical JPEG compression method.



a)Input Image



b)Segmented Main Subject Area

Figure - 7 Input /Output

Table -1 Compression Ratio and PSNR values obtained by Hybrid JPEG and JPEG

Image		LQ	Ну	/brid	Hybrid/HQ	I	HQ	
	CR1	PSNR1	CR2	PSNR2	PSNR@	CR3	PSNR3	CR2-CR3
					MainSubject			
3	26.00	21.52	7.46	27.82	27.84	7.46	27.82	0.0000
5	25.29	22.09	6.80	28.87	28.93	6.80	28.87	0.0004
10	23.83	20.78	5.41	28.09	28.13	5.41	28.09	0.0031
4	24.33	22.08	6.61	31.13	31.20	6.60	31.13	0.0068
11	27.75	25.33	8.62	35.97	36.25	8.61	36.01	0.0198
9	24.92	21.62	6.58	30.45	30.56	6.56	30.51	0.0204
1	27.54	23.03	8.40	29.37	29.40	8.37	29.37	0.0276
7	24.85	17.71	4.01	23.38	23.43	3.92	23.38	0.0911
8	24.63	21.97	6.73	31.05	31.13	6.64	31.10	0.0921
12	29.18	22.73	5.92	28.70	29.07	5.61	28.93	0.3111
6	26.47	20.12	5.54	25.31	25.33	5.19	25.30	0.3449
2	22.84	20.40	7.91	27.93	28.21	7.25	28.06	0.6541



Figure – 8 Test Images (1-12 from Left to Right and Top to Bottom)

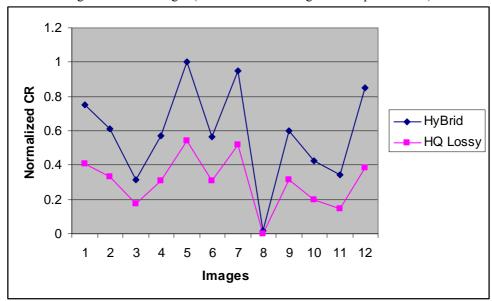


Figure – 9 Normalized Compression Ratio Obtained for Test Images

Increased Compression Ratio by Hybrid JPEG Compression

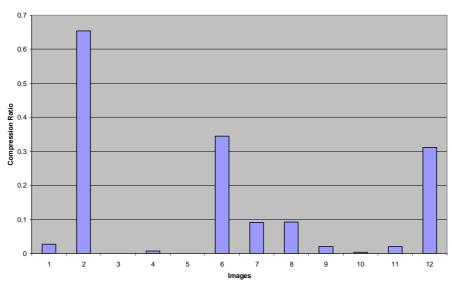


Figure – 10 Increased Compression Ratios by Hybrid Compression.

V. CONCLUSION

The compression ratio of Hybrid JPEG method is higher than JPEG method in more than 90% of test cases. In the worst case both Hybrid JPEG and JPEG method gives the same compression ratio. The PSNR value at the main subject area is same for both methods. The PSNR value at

the background area is lower in Hybrid JPEG method which is acceptable, since the background area is not vital. The Hybrid JPEG method is suitable for imagery with larger trivial background and certain level of loss is permissible.

ACKNOWLEDGEMENTS

The authors express their gratitude to University Grant Commission and Manonmanium Sundaranar University for financial assistance under the Faculty Development Program.

REFERENCES

- [1] Xiwen OwenZhao, Zhihai HenryHe, "Lossless Image Compression Using Super-Spatial Structure Prediction", IEEE Signal Processing Letters, vol. 17, no. 4, April 2010 383
- [2] Aaron T. Deever and Sheila S. Hemami, "Lossless Image Compression With Projection-Based and Adaptive Reversible Integer Wavelet Transforms", *IEEE Transactions on Image Processing*, vol. 12, NO. 5, May 2003.
- [3] Nikolaos V. Boulgouris, Dimitrios Tzovaras, and Michael Gerassimos Strintzis, "Lossless Image Compression Based on OptimalPrediction, Adaptive Lifting, and Conditional Arithmetic Coding", IEEE Transactions on Image Processing, vol. 10, NO. 1, Jan 2001
- [4] Xin Li, , and Michael T. Orchard "Edge-Directed Prediction for Lossless Compression of Natural Images", *IEEE Transactions on Image Processing*, vol. 10, NO. 6, Jun 2001.
- [5] Jaemoon Kim, Jungsoo Kim and Chong-Min Kyung, "A Lossless Embedded Compression Algorithm for High Definition Video Coding" 978-1-4244-4291 / 09 2009 IEEE. ICME 2009
- [6] Rene J. van der Vleuten, Richard P.Kleihorstt ,Christian Hentschel,t "Low-Complexity Scalable DCT Image Compression", 2000 IEEE.
- [7] K.Somasundaram, and S.Domnic, "Modified Vector Quantization Method for mage Compression", "Transactions On Engineering, Computing And Technology Vol 13 May 2006
- [8] Mohamed A. El-Sharkawy, Chstian A. White and Harry ,"Subband Image Compression Using Wavelet Transform And Vector Quantization", 1997 IEEE.
- [9] Amir Averbuch, Danny Lazar, and Moshe Israeli ,"Image Compression Using Wavelet Transform and Multiresolution Decomposition", *IEEE Transactions on Image Processing*, vol. 5, NO. 1, Jan 1996.
- [10] Jianyu Lin, Mark J. T. Smith," New Perspectives and Improvements on the Symmetric Extension Filter Bank for Subband /Wavelet Image Compression", IEEE Transactions on Image Processing, vol. 17, NO. 2, Feb 2008.
- [11] Yu Liu, Student Member, and King Ngi Ngan, "Weighted Adaptive Lifting-Based Wavelet Transform for Image Coding ",IEEE Transactions on Image Processing, vol. 17, NO. 4, Apr 2008.
- [12] Michael B. Martin and Amy E. Bell, "New Image Compression Techniques Using Multiwavelets and Multiwavelet Packets" "IEEE Transactions on Image Processing, vol. 10, NO. 4, Apr 2001
- [13] Roger L. Claypoole, Jr , Geoffrey M. Davis, Wim Sweldens ,"Nonlinear Wavelet Transforms for Image Coding via Lifting", IEEE Transactions on Image Processing, vol. 12, NO. 12, Dec 2003
- [14] David Salomon, "Data Compression, Complete Reference", Springer-Verlag New York, Inc, ISBN 0-387-40697-2.
- [15] Eddie Batista de Lima Filho, Eduardo A. B. da Silva Murilo Bresciani de Carvalho, and Frederico Silva Pinagé "Universal Image Compression Using Multiscale Recurrent Patterns With Adaptive Probability Model", IEEE Transactions on Image Processing, vol. 17, NO. 4, Apr 2008.
- [16] Ingo Bauermann, and Eckehard Steinbach, "RDTC Optimized Compression of Image-Based Scene Representations (Part I): Modeling and Theoretical Analysis", IEEE Transactions on Image Processing, vol. 17, NO. 5, May 2008.
- [17] Roman Kazinnik, Shai Dekel, and Nira Dyn, "Low Bit-Rate Image Coding Using Adaptive Geometric Piecewise Polynomial Approximation", *IEEE Transactions on Image Processing*, vol. 16, NO. 9, Sep 2007.
- [18] Marta Mrak, Sonja Grgic, and Mislav Grgic, "Picture Quality Measures in Image Compression Systems", EUROCON 2003 Ljubljana, Slovenia, 0-7803-7763-W03 2003 IEEE.
- [19] Alan C. Brooks, Xiaonan Zhao, , Thrasyvoulos N. Pappas., "Structural Similarity Quality Metrics in a Coding Context:

- Exploring the Space of Realistic Distortions", *IEEE Transactions on Image Processing*, vol. 17, no. 8, pp. 1261–1273, Aug 2008.
- [20] Hong, S. W. Bao, P., "Hybrid image compression model based on subband coding and edge-preserving regularization", Vision, Image and Signal Processing, IEE Proceedings, Volume: 147, Issue: 1, 16-22, Feb 2000
- [21] Zhe-Ming Lu, Hui Pei ,"Hybrid Image Compression Scheme Based on PVQ and DCTVQ ",IEICE - Transactions on Information and Systems archive, Vol E88-D, Issue 10 ,October 2006
- [22] Y.Jacob Vetha Raj, M.Mohamed Sathik and K.Senthamarai Kanna,, "Hybrid Image Compression by Blurring Background and Non-Edges. The International Journal on Multimedia and its applications. Vol 2, No. 1, February 2010, pp 32-41
- [23] Willian K. Pratt ,"Digital Image Processing" ,John Wiley & Sons, Inc, ISBN 9-814-12620-9.
- [24] Jundi Ding, Runing Ma, and Songcan Chen,"A Scale-Based Connected Coherence Tree Algorithm for Image Segmentation", *IEEE Transactions on Image Processing*, vol. 17, NO. 2, Feb 2008
- [25] Kyungsuk (Peter) Pyun, , Johan Lim, Chee Sun Won, and Robert M. Gray, "Image Segmentation Using Hidden Markov Gauss Mixture Models", " JEEE Transactions on Image Processing, VOL. 16, NO. 7, JULY 2007

IJCSIS REVIEWERS' LIST

Assist Prof (Dr.) M. Emre Celebi, Louisiana State University in Shreveport, USA

Dr. Lam Hong Lee, Universiti Tunku Abdul Rahman, Malaysia

Dr. Shimon K. Modi, Director of Research BSPA Labs, Purdue University, USA

Dr. Jianguo Ding, Norwegian University of Science and Technology (NTNU), Norway

Assoc. Prof. N. Jaisankar, VIT University, Vellore, Tamilnadu, India

Dr. Amogh Kavimandan, The Mathworks Inc., USA

Dr. Ramasamy Mariappan, Vinayaka Missions University, India

Dr. Yong Li, School of Electronic and Information Engineering, Beijing Jiaotong University, P.R. China

Assist. Prof. Sugam Sharma, NIET, India / Iowa State University, USA

Dr. Jorge A. Ruiz-Vanoye, Universidad Autónoma del Estado de Morelos, Mexico

Dr. Neeraj Kumar, SMVD University, Katra (J&K), India

Dr Genge Bela, "Petru Maior" University of Targu Mures, Romania

Dr. Junjie Peng, Shanghai University, P. R. China

Dr. Ilhem LENGLIZ, HANA Group - CRISTAL Laboratory, Tunisia

Prof. Dr. Durgesh Kumar Mishra, Acropolis Institute of Technology and Research, Indore, MP, India

Jorge L. Hernández-Ardieta, University Carlos III of Madrid, Spain

Prof. Dr.C.Suresh Gnana Dhas, Anna University, India

Mrs Li Fang, Nanyang Technological University, Singapore

Prof. Pijush Biswas, RCC Institute of Information Technology, India

Dr. Siddhivinayak Kulkarni, University of Ballarat, Ballarat, Victoria, Australia

Dr. A. Arul Lawrence, Royal College of Engineering & Technology, India

Mr. Wongyos Keardsri, Chulalongkorn University, Bangkok, Thailand

Mr. Somesh Kumar Dewangan, CSVTU Bhilai (C.G.)/ Dimat Raipur, India

Mr. Hayder N. Jasem, University Putra Malaysia, Malaysia

Mr. A.V.Senthil Kumar, C. M. S. College of Science and Commerce, India

Mr. R. S. Karthik, C. M. S. College of Science and Commerce, India

Mr. P. Vasant, University Technology Petronas, Malaysia

Mr. Wong Kok Seng, Soongsil University, Seoul, South Korea

Mr. Praveen Ranjan Srivastava, BITS PILANI, India

Mr. Kong Sang Kelvin, Leong, The Hong Kong Polytechnic University, Hong Kong

Mr. Mohd Nazri Ismail, Universiti Kuala Lumpur, Malaysia

Dr. Rami J. Matarneh, Al-isra Private University, Amman, Jordan

Dr Ojesanmi Olusegun Ayodeji, Ajayi Crowther University, Oyo, Nigeria

Dr. Riktesh Srivastava, Skyline University, UAE

Dr. Oras F. Baker, UCSI University - Kuala Lumpur, Malaysia

Dr. Ahmed S. Ghiduk, Faculty of Science, Beni-Suef University, Egypt

and Department of Computer science, Taif University, Saudi Arabia

Mr. Tirthankar Gayen, IIT Kharagpur, India

Ms. Huei-Ru Tseng, National Chiao Tung University, Taiwan

Prof. Ning Xu, Wuhan University of Technology, China

Mr Mohammed Salem Binwahlan, Hadhramout University of Science and Technology, Yemen

& Universiti Teknologi Malaysia, Malaysia.

Dr. Aruna Ranganath, Bhoj Reddy Engineering College for Women, India

Mr. Hafeezullah Amin, Institute of Information Technology, KUST, Kohat, Pakistan

Prof. Syed S. Rizvi, University of Bridgeport, USA

Mr. Shahbaz Pervez Chattha, University of Engineering and Technology Taxila, Pakistan

Dr. Shishir Kumar, Jaypee University of Information Technology, Wakanaghat (HP), India

Mr. Shahid Mumtaz, Portugal Telecommunication, Instituto de Telecomunicações (IT), Aveiro, Portugal

Mr. Rajesh K Shukla, Corporate Institute of Science & Technology Bhopal M P

Dr. Poonam Garg, Institute of Management Technology, India

Mr. S. Mehta, Inha University, Korea

Mr. Dilip Kumar S.M, University Visvesvaraya College of Engineering (UVCE), Bangalore University, Bangalore

Prof. Malik Sikander Hayat Khiyal, Fatima Jinnah Women University, Rawalpindi, Pakistan

Dr. Virendra Gomase, Department of Bioinformatics, Padmashree Dr. D.Y. Patil University

Dr. Irraivan Elamvazuthi, University Technology PETRONAS, Malaysia

Mr. Saqib Saeed, University of Siegen, Germany

Mr. Pavan Kumar Gorakavi, IPMA-USA [YC]

Dr. Ahmed Nabih Zaki Rashed, Menoufia University, Egypt

Prof. Shishir K. Shandilya, Rukmani Devi Institute of Science & Technology, India

Mrs.J.Komala Lakshmi, SNR Sons College, Computer Science, India

Mr. Muhammad Sohail, KUST, Pakistan

Dr. Manjaiah D.H, Mangalore University, India

Dr. S Santhosh Baboo, D.G.Vaishnav College, Chennai, India

Prof. Dr. Mokhtar Beldjehem, Sainte-Anne University, Halifax, NS, Canada

Dr. Deepak Laxmi Narasimha, Faculty of Computer Science and Information Technology, University of Malaya, Malaysia

Prof. Dr. Arunkumar Thangavelu, Vellore Institute Of Technology, India

Mr. M. Azath, Anna University, India

Mr. Md. Rabiul Islam, Rajshahi University of Engineering & Technology (RUET), Bangladesh

Mr. Aos Alaa Zaidan Ansaef, Multimedia University, Malaysia

Dr Suresh Jain, Professor (on leave), Institute of Engineering & Technology, Devi Ahilya University, Indore (MP) India,

Dr. Mohammed M. Kadhum, Universiti Utara Malaysia

Mr. Hanumanthappa. J. University of Mysore, India

Mr. Syed Ishtiaque Ahmed, Bangladesh University of Engineering and Technology (BUET)

Mr Akinola Solomon Olalekan, University of Ibadan, Ibadan, Nigeria

Mr. Santosh K. Pandey, Department of Information Technology, The Institute of Chartered Accountants of India

Dr. P. Vasant, Power Control Optimization, Malaysia

Dr. Petr Ivankov, Automatika - S, Russian Federation

Dr. Utkarsh Seetha, Data Infosys Limited, India

Mrs. Priti Maheshwary, Maulana Azad National Institute of Technology, Bhopal

Dr. (Mrs) Padmavathi Ganapathi, Avinashilingam University for Women, Coimbatore

Assist. Prof. A. Neela madheswari, Anna university, India

Prof. Ganesan Ramachandra Rao, PSG College of Arts and Science, India

Mr. Kamanashis Biswas, Daffodil International University, Bangladesh

Dr. Atul Gonsai, Saurashtra University, Gujarat, India

Mr. Angkoon Phinyomark, Prince of Songkla University, Thailand

Mrs. G. Nalini Priya, Anna University, Chennai

Dr. P. Subashini, Avinashilingam University for Women, India

Assoc. Prof. Vijay Kumar Chakka, Dhirubhai Ambani IICT, Gandhinagar ,Gujarat

Mr Jitendra Agrawal, : Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal

Mr. Vishal Goyal, Department of Computer Science, Punjabi University, India

Dr. R. Baskaran, Department of Computer Science and Engineering, Anna University, Chennai

Assist. Prof, Kanwalvir Singh Dhindsa, B.B.S.B.Engg.College, Fatehgarh Sahib (Punjab), India

Dr. Jamal Ahmad Dargham, School of Engineering and Information Technology, Universiti Malaysia Sabah

Mr. Nitin Bhatia, DAV College, India

Dr. Dhavachelvan Ponnurangam, Pondicherry Central University, India

Dr. Mohd Faizal Abdollah, University of Technical Malaysia, Malaysia

Assist. Prof. Sonal Chawla, Panjab University, India

Dr. Abdul Wahid, AKG Engg. College, Ghaziabad, India

Mr. Arash Habibi Lashkari, University of Malaya (UM), Malaysia

Mr. Md. Rajibul Islam, Ibnu Sina Institute, University Technology Malaysia

Professor Dr. Sabu M. Thampi, .B.S Institute of Technology for Women, Kerala University, India

Mr. Noor Muhammed Nayeem, Université Lumière Lyon 2, 69007 Lyon, France

Dr. Himanshu Aggarwal, Department of Computer Engineering, Punjabi University, India

Prof R. Naidoo, Dept of Mathematics/Center for Advanced Computer Modelling, Durban University of Technology, Durban, South Africa

Prof. Mydhili K Nair, M S Ramaiah Institute of Technology(M.S.R.I.T), Affliliated to Visweswaraiah Technological University, Bangalore, India

M. Prabu, Adhiyamaan College of Engineering/Anna University, India

Mr. Swakkhar Shatabda, Department of Computer Science and Engineering, United International University, Bangladesh

Dr. Abdur Rashid Khan, ICIT, Gomal University, Dera Ismail Khan, Pakistan

Mr. H. Abdul Shabeer, I-Nautix Technologies, Chennai, India

Dr. M. Aramudhan, Perunthalaivar Kamarajar Institute of Engineering and Technology, India

Dr. M. P. Thapliyal, Department of Computer Science, HNB Garhwal University (Central University), India

Dr. Shahaboddin Shamshirband, Islamic Azad University, Iran

Mr. Zeashan Hameed Khan, : Université de Grenoble, France

Prof. Anil K Ahlawat, Ajay Kumar Garg Engineering College, Ghaziabad, UP Technical University, Lucknow

Mr. Longe Olumide Babatope, University Of Ibadan, Nigeria

Associate Prof. Raman Maini, University College of Engineering, Punjabi University, India

Dr. Maslin Masrom, University Technology Malaysia, Malaysia

Sudipta Chattopadhyay, Jadavpur University, Kolkata, India

Dr. Dang Tuan NGUYEN, University of Information Technology, Vietnam National University - Ho Chi Minh City

Dr. Mary Lourde R., BITS-PILANI Dubai, UAE

Dr. Abdul Aziz, University of Central Punjab, Pakistan

Mr. Karan Singh, Gautam Budtha University, India

Mr. Avinash Pokhriyal, Uttar Pradesh Technical University, Lucknow, India

Associate Prof Dr Zuraini Ismail, University Technology Malaysia, Malaysia

Assistant Prof. Yasser M. Alginahi, College of Computer Science and Engineering, Taibah University, Madinah Munawwarrah, KSA

Mr. Dakshina Ranjan Kisku, West Bengal University of Technology, India

Mr. Raman Kumar, Dr B R Ambedkar National Institute of Technology, Jalandhar, Punjab, India

Associate Prof. Samir B. Patel, Institute of Technology, Nirma University, India

Dr. M.Munir Ahamed Rabbani, B. S. Abdur Rahman University, India

Asst. Prof. Koushik Majumder, West Bengal University of Technology, India

Dr. Alex Pappachen James, Queensland Micro-nanotechnology center, Griffith University, Australia

Assistant Prof. S. Hariharan, B.S. Abdur Rahman University, India

Asst Prof. Jasmine. K. S, R.V.College of Engineering, India

Mr Naushad Ali Mamode Khan, Ministry of Education and Human Resources, Mauritius

Prof. Mahesh Goyani, G H Patel Collge of Engg. & Tech, V.V.N, Anand, Gujarat, India

Dr. Mana Mohammed, University of Tlemcen, Algeria

Prof. Jatinder Singh, Universal Institution of Engg. & Tech. CHD, India

Mrs. M. Anandhavalli Gauthaman, Sikkim Manipal Institute of Technology, Majitar, East Sikkim

Dr. Bin Guo, Institute Telecom SudParis, France

Mrs. Maleika Mehr Nigar Mohamed Heenaye-Mamode Khan, University of Mauritius

Prof. Pijush Biswas, RCC Institute of Information Technology, India

Mr. V. Bala Dhandayuthapani, Mekelle University, Ethiopia

Mr. Irfan Syamsuddin, State Polytechnic of Ujung Pandang, Indonesia

Mr. Kavi Kumar Khedo, University of Mauritius, Mauritius

Mr. Ravi Chandiran, Zagro Singapore Pte Ltd. Singapore

Mr. Milindkumar V. Sarode, Jawaharlal Darda Institute of Engineering and Technology, India

Dr. Shamimul Qamar, KSJ Institute of Engineering & Technology, India

Dr. C. Arun, Anna University, India

Assist. Prof. M.N.Birje, Basaveshwar Engineering College, India

Prof. Hamid Reza Naji, Department of Computer Enigneering, Shahid Beheshti University, Tehran, Iran

Assist. Prof. Debasis Giri, Department of Computer Science and Engineering, Haldia Institute of Technology Subhabrata Barman, Haldia Institute of Technology, West Bengal

Mr. M. I. Lali, COMSATS Institute of Information Technology, Islamabad, Pakistan

Dr. Feroz Khan, Central Institute of Medicinal and Aromatic Plants, Lucknow, India

Mr. R. Nagendran, Institute of Technology, Coimbatore, Tamilnadu, India

Mr. Amnach Khawne, King Mongkut's Institute of Technology Ladkrabang, Ladkrabang, Bangkok, Thailand

Dr. P. Chakrabarti, Sir Padampat Singhania University, Udaipur, India

Mr. Nafiz Imtiaz Bin Hamid, Islamic University of Technology (IUT), Bangladesh.

Shahab-A. Shamshirband, Islamic Azad University, Chalous, Iran

Prof. B. Priestly Shan, Anna Univeristy, Tamilnadu, India

Venkatramreddy Velma, Dept. of Bioinformatics, University of Mississippi Medical Center, Jackson MS USA Akshi Kumar, Dept. of Computer Engineering, Delhi Technological University, India

Dr. Umesh Kumar Singh, Vikram University, Ujjain, India

Mr. Serguei A. Mokhov, Concordia University, Canada

Mr. Lai Khin Wee, Universiti Teknologi Malaysia, Malaysia

Dr. Awadhesh Kumar Sharma, Madan Mohan Malviya Engineering College, India

Mr. Syed R. Rizvi, Analytical Services & Materials, Inc., USA

Dr. S. Karthik, SNS Collegeof Technology, India

Mr. Syed Qasim Bukhari, CIMET (Universidad de Granada), Spain

Mr. A.D.Potgantwar, Pune University, India

Dr. Himanshu Aggarwal, Punjabi University, India

Mr. Rajesh Ramachandran, Naipunya Institute of Management and Information Technology, India

Dr. K.L. Shunmuganathan, R.M.K Engg College , Kavaraipettai ,Chennai

Dr. Prasant Kumar Pattnaik, KIST, India.

Dr. Ch. Aswani Kumar, VIT University, India

Mr. Ijaz Ali Shoukat, King Saud University, Riyadh KSA

Mr. Arun Kumar, Sir Padam Pat Singhania University, Udaipur, Rajasthan

Mr. Muhammad Imran Khan, Universiti Teknologi PETRONAS, Malaysia

Dr. Natarajan Meghanathan, Jackson State University, Jackson, MS, USA

Mr. Mohd Zaki Bin Mas'ud, Universiti Teknikal Malaysia Melaka (UTeM), Malaysia

Prof. Dr. R. Geetharamani, Dept. of Computer Science and Eng., Rajalakshmi Engineering College, India

Dr. Smita Rajpal, Institute of Technology and Management, Gurgaon, India

Dr. S. Abdul Khader Jilani, University of Tabuk, Tabuk, Saudi Arabia

Mr. Syed Jamal Haider Zaidi, Bahria University, Pakistan

Dr. N. Devarajan, Government College of Technology, Coimbatore, Tamilnadu, INDIA

Mr. R. Jagadeesh Kannan, RMK Engineering College, India

Mr. Deo Prakash, Shri Mata Vaishno Devi University, India

Mr. Mohammad Abu Naser, Dept. of EEE, IUT, Gazipur, Bangladesh

Assist. Prof. Prasun Ghosal, Bengal Engineering and Science University, India

Mr. Md. Golam Kaosar, School of Engineering and Science, Victoria University, Melbourne City, Australia

Mr. R. Mahammad Shafi, Madanapalle Institute of Technology & Science, India

Dr. F.Sagayaraj Francis, Pondicherry Engineering College, India

Dr. Ajay Goel, HIET, Kaithal, India

Mr. Nayak Sunil Kashibarao, Bahirji Smarak Mahavidyalaya, India

Mr. Suhas J Manangi, Microsoft India

Dr. Kalyankar N. V., Yeshwant Mahavidyalaya, Nanded, India

Dr. K.D. Verma, S.V. College of Post graduate studies & Research, India

Dr. Amjad Rehman, University Technology Malaysia, Malaysia

Mr. Rachit Garg, L K College, Jalandhar, Punjab

Mr. J. William, M.A.M college of Engineering, Trichy, Tamilnadu, India

Prof. Jue-Sam Chou, Nanhua University, College of Science and Technology, Taiwan

Dr. Thorat S.B., Institute of Technology and Management, India

Mr. Ajay Prasad, Sir Padampat Singhania University, Udaipur, India

Dr. Kamaljit I. Lakhtaria, Atmiya Institute of Technology & Science, India

Mr. Syed Rafiul Hussain, Ahsanullah University of Science and Technology, Bangladesh

Mrs Fazeela Tunnisa, Najran University, Kingdom of Saudi Arabia

Mrs Kavita Taneja, Maharishi Markandeshwar University, Haryana, India

Mr. Maniyar Shiraz Ahmed, Najran University, Najran, KSA

Mr. Anand Kumar, AMC Engineering College, Bangalore

Dr. Rakesh Chandra Gangwar, Beant College of Engg. & Tech., Gurdaspur (Punjab) India

Dr. V V Rama Prasad, Sree Vidyanikethan Engineering College, India

Assist. Prof. Neetesh Kumar Gupta, Technocrats Institute of Technology, Bhopal (M.P.), India

Mr. Ashish Seth, Uttar Pradesh Technical University, Lucknow ,UP India

Dr. V V S S S Balaram, Sreenidhi Institute of Science and Technology, India

Mr Rahul Bhatia, Lingaya's Institute of Management and Technology, India

Prof. Niranjan Reddy. P, KITS, Warangal, India

Prof. Rakesh. Lingappa, Vijetha Institute of Technology, Bangalore, India

Dr. Mohammed Ali Hussain, Nimra College of Engineering & Technology, Vijayawada, A.P., India

Dr. A.Srinivasan, MNM Jain Engineering College, Rajiv Gandhi Salai, Thorapakkam, Chennai

Mr. Rakesh Kumar, M.M. University, Mullana, Ambala, India

Dr. Lena Khaled, Zarqa Private University, Aman, Jordon

Ms. Supriya Kapoor, Patni/Lingaya's Institute of Management and Tech., India

Dr. Tossapon Boongoen, Aberystwyth University, UK

Dr. Bilal Alatas, Firat University, Turkey

Assist. Prof. Jyoti Praaksh Singh, Academy of Technology, India

Dr. Ritu Soni, GNG College, India

Dr . Mahendra Kumar , Sagar Institute of Research & Technology, Bhopal, India.

Dr. Binod Kumar, India

Dr. Muzhir Shaban Al-Ani, Amman Arab University Amman - Jordan

Dr. T.C. Manjunath, ATRIA Institute of Tech, India

Mr. Muhammad Zakarya, COMSATS Institute of Information Technology (CIIT), Pakistan

Assist. Prof. Harmunish Taneja, M. M. University, India

Dr. Chitra Dhawale, SICSR, Model Colony, Pune, India

Mrs Sankari Muthukaruppan, Nehru Institute of Engineering and Technology, Anna University, India

Mr. Aaqif Afzaal Abbasi, National University Of Sciences And Technology, Islamabad

Prof. Ashutosh Kumar Dubey, Trinity Institute of Technology and Research Bhopal, India

Mr. G. Appasami, Dr. Pauls Engineering College, India

Mr. M Yasin, National University of Science and Tech, karachi (NUST), Pakistan

Mr. Yaser Miaji, University Utara Malaysia, Malaysia

Mr. Shah Ahsanul Haque, International Islamic University Chittagong (IIUC), Bangladesh

Prof. (Dr) Syed Abdul Sattar, Royal Institute of Technology & Science, India

Dr. S. Sasikumar, Roever Engineering College

Assist. Prof. Monit Kapoor, Maharishi Markandeshwar University, India

Mr. Nwaocha Vivian O, National Open University of Nigeria

Dr. M. S. Vijaya, GR Govindarajulu School of Applied Computer Technology, India

Assist. Prof. Chakresh Kumar, Manav Rachna International University, India

Mr. Kunal Chadha, R&D Software Engineer, Gemalto, Singapore

Mr. Pawan Jindal, Jaypee University of Engineering and Technology, India

Mr. Mueen Uddin, Universiti Teknologi Malaysia, UTM, Malaysia

Dr. Dhuha Basheer abdullah, Mosul university, Iraq

Mr. S. Audithan, Annamalai University, India

Prof. Vijay K Chaudhari, Technocrats Institute of Technology, India

Associate Prof. Mohd Ilyas Khan, Technocrats Institute of Technology, India

Dr. Vu Thanh Nguyen, University of Information Technology, HoChiMinh City, VietNam

Assist. Prof. Anand Sharma, MITS, Lakshmangarh, Sikar, Rajasthan, India

Prof. T V Narayana Rao, HITAM Engineering college, Hyderabad

Mr. Deepak Gour, Sir Padampat Singhania University, India

Assist. Prof. Amutharaj Joyson, Kalasalingam University, India

Mr. Ali Balador, Islamic Azad University, Iran

Mr. Mohit Jain, Maharaja Surajmal Institute of Technology, India

Mr. Dilip Kumar Sharma, GLA Institute of Technology & Management, India

Dr. Debojyoti Mitra, Sir padampat Singhania University, India

Dr. Ali Dehghantanha, Asia-Pacific University College of Technology and Innovation, Malaysia

Mr. Zhao Zhang, City University of Hong Kong, China

Prof. S.P. Setty, A.U. College of Engineering, India

Prof. Patel Rakeshkumar Kantilal, Sankalchand Patel College of Engineering, India

Mr. Biswajit Bhowmik, Bengal College of Engineering & Technology, India

Mr. Manoj Gupta, Apex Institute of Engineering & Technology, India

Assist. Prof. Ajay Sharma, Raj Kumar Goel Institute Of Technology, India

Assist. Prof. Ramveer Singh, Raj Kumar Goel Institute of Technology, India

Dr. Hanan Elazhary, Electronics Research Institute, Egypt

Dr. Hosam I. Faiq, USM, Malaysia

Prof. Dipti D. Patil, MAEER's MIT College of Engg. & Tech, Pune, India

Assist. Prof. Devendra Chack, BCT Kumaon engineering College Dwarahat Almora, India

Prof. Manpreet Singh, M. M. Engg. College, M. M. University, India

Assist. Prof. M. Sadiq ali Khan, University of Karachi, Pakistan

Mr. Prasad S. Halgaonkar, MIT - College of Engineering, Pune, India

Dr. Imran Ghani, Universiti Teknologi Malaysia, Malaysia

Prof. Varun Kumar Kakar, Kumaon Engineering College, Dwarahat, India

Assist. Prof. Nisheeth Joshi, Apaji Institute, Banasthali University, Rajasthan, India

Associate Prof. Kunwar S. Vaisla, VCT Kumaon Engineering College, India

Prof Anupam Choudhary, Bhilai School Of Engg., Bhilai (C.G.), India

Mr. Divya Prakash Shrivastava, Al Jabal Al garbi University, Zawya, Libya

Associate Prof. Dr. V. Radha, Avinashilingam Deemed university for women, Coimbatore.

Dr. Kasarapu Ramani, JNT University, Anantapur, India

Dr. Anuraag Awasthi, Jayoti Vidyapeeth Womens University, India

Dr. C G Ravichandran, R V S College of Engineering and Technology, India

Dr. Mohamed A. Deriche, King Fahd University of Petroleum and Minerals, Saudi Arabia

Mr. Abbas Karimi, Universiti Putra Malaysia, Malaysia

Mr. Amit Kumar, Jaypee University of Engg. and Tech., India

Dr. Nikolai Stoianov, Defense Institute, Bulgaria

Assist. Prof. S. Ranichandra, KSR College of Arts and Science, Tiruchencode

Mr. T.K.P. Rajagopal, Diamond Horse International Pvt Ltd, India

Dr. Md. Ekramul Hamid, Rajshahi University, Bangladesh

Mr. Hemanta Kumar Kalita, TATA Consultancy Services (TCS), India

Dr. Messaouda Azzouzi, Ziane Achour University of Djelfa, Algeria

Prof. (Dr.) Juan Jose Martinez Castillo, "Gran Mariscal de Ayacucho" University and Acantelys research Group, Venezuela

Dr. Jatinderkumar R. Saini, Narmada College of Computer Application, India

Dr. Babak Bashari Rad, University Technology of Malaysia, Malaysia

Mr. B. Muthu Kumar, Kathir College Of Engineering, Coimbatore

Dr. Nighat Mir, Effat University, Saudi Arabia

Prof. (Dr.) G.M.Nasira, Sasurie College of Engineering, India

Mr. Varun Mittal, Gemalto Pte Ltd, Singapore

Assist. Prof. Mrs P. Banumathi, Kathir College Of Engineering, Coimbatore

Assist. Prof. Quan Yuan, University of Wisconsin-Stevens Point, US

Dr. Pranam Paul, Narula Institute of Technology, Agarpara, West Bengal, India

Assist. Prof. J. Ramkumar, V.L.B Janakiammal college of Arts & Science, India

Mr. P. Sivakumar, Anna university, Chennai, India

Mr. Md. Humayun Kabir Biswas, King Khalid University, Kingdom of Saudi Arabia

Mr. Mayank Singh, J.P. Institute of Engg & Technology, Meerut, India

HJ. Kamaruzaman Jusoff, Universiti Putra Malaysia

Mr. Nikhil Patrick Lobo, CADES, India

Mr. Amit Wason, Rayat-Bahra Institute of Engineering & Boi-Technology, India

Dr. Rajesh Shrivastava, Govt. Benazir Science & Commerce College, Bhopal, India

CALL FOR PAPERS

International Journal of Computer Science and Information Security IJCSIS 2011

ISSN: 1947-5500

http://sites.google.com/site/ijcsis/

International Journal Computer Science and Information Security, IJCSIS, is the premier scholarly venue in the areas of computer science and security issues. IJCSIS 2011 will provide a high profile, leading edge platform for researchers and engineers alike to publish state-of-the-art research in the respective fields of information technology and communication security. The journal will feature a diverse mixture of publication articles including core and applied computer science related topics.

Authors are solicited to contribute to the special issue by submitting articles that illustrate research results, projects, surveying works and industrial experiences that describe significant advances in the following areas, but are not limited to. Submissions may span a broad range of topics, e.g.:

Track A: Security

Access control, Anonymity, Audit and audit reduction & Authentication and authorization, Applied cryptography, Cryptanalysis, Digital Signatures, Biometric security, Boundary control devices, Certification and accreditation, Cross-layer design for security, Security & Network Management, Data and system integrity, Database security, Defensive information warfare, Denial of service protection, Intrusion Detection, Anti-malware, Distributed systems security, Electronic commerce, E-mail security, Spam, Phishing, E-mail fraud, Virus, worms, Trojan Protection, Grid security, Information hiding and watermarking & Information survivability, Insider threat protection, Integrity

Intellectual property protection, Internet/Intranet Security, Key management and key recovery, Languagebased security, Mobile and wireless security, Mobile, Ad Hoc and Sensor Network Security, Monitoring and surveillance, Multimedia security, Operating system security, Peer-to-peer security, Performance Evaluations of Protocols & Security Application, Privacy and data protection, Product evaluation criteria and compliance, Risk evaluation and security certification, Risk/vulnerability assessment, Security & Network Management, Security Models & protocols, Security threats & countermeasures (DDoS, MiM, Session Hijacking, Replay attack etc.), Trusted computing, Ubiquitous Computing Security, Virtualization security, VoIP security, Web 2.0 security, Submission Procedures, Active Defense Systems, Adaptive Defense Systems, Benchmark, Analysis and Evaluation of Security Systems, Distributed Access Control and Trust Management, Distributed Attack Systems and Mechanisms, Distributed Intrusion Detection/Prevention Systems, Denial-of-Service Attacks and Countermeasures, High Performance Security Systems, Identity Management and Authentication, Implementation, Deployment and Management of Security Systems, Intelligent Defense Systems, Internet and Network Forensics, Largescale Attacks and Defense, RFID Security and Privacy, Security Architectures in Distributed Network Systems, Security for Critical Infrastructures, Security for P2P systems and Grid Systems, Security in E-Commerce, Security and Privacy in Wireless Networks, Secure Mobile Agents and Mobile Code, Security Protocols, Security Simulation and Tools, Security Theory and Tools, Standards and Assurance Methods, Trusted Computing, Viruses, Worms, and Other Malicious Code, World Wide Web Security, Novel and emerging secure architecture, Study of attack strategies, attack modeling, Case studies and analysis of actual attacks, Continuity of Operations during an attack, Key management, Trust management, Intrusion detection techniques, Intrusion response, alarm management, and correlation analysis, Study of tradeoffs between security and system performance, Intrusion tolerance systems, Secure protocols, Security in wireless networks (e.g. mesh networks, sensor networks, etc.), Cryptography and Secure Communications, Computer Forensics, Recovery and Healing, Security Visualization, Formal Methods in Security, Principles for Designing a Secure Computing System, Autonomic Security, Internet Security, Security in Health Care Systems, Security Solutions Using Reconfigurable Computing, Adaptive and Intelligent Defense Systems, Authentication and Access control, Denial of service attacks and countermeasures, Identity, Route and Location Anonymity schemes, Intrusion detection and prevention techniques, Cryptography, encryption algorithms and Key management schemes, Secure routing schemes, Secure neighbor discovery and localization, Trust establishment and maintenance, Confidentiality and data integrity, Security architectures, deployments and solutions, Emerging threats to cloud-based services, Security model for new services, Cloud-aware web service security, Information hiding in Cloud Computing, Securing distributed data storage in cloud, Security, privacy and trust in mobile computing systems and applications, **Middleware security & Security features:** middleware software is an asset on

its own and has to be protected, interaction between security-specific and other middleware features, e.g., context-awareness, Middleware-level security monitoring and measurement: metrics and mechanisms for quantification and evaluation of security enforced by the middleware, Security co-design: trade-off and co-design between application-based and middleware-based security, Policy-based management: innovative support for policy-based definition and enforcement of security concerns, Identification and authentication mechanisms: Means to capture application specific constraints in defining and enforcing access control rules, Middleware-oriented security patterns: identification of patterns for sound, reusable security, Security in aspect-based middleware: mechanisms for isolating and enforcing security aspects, Security in agent-based platforms: protection for mobile code and platforms, Smart Devices: Biometrics, National ID cards, Embedded Systems Security and TPMs, RFID Systems Security, Smart Card Security, Pervasive Systems: Digital Rights Management (DRM) in pervasive environments, Intrusion Detection and Information Filtering, Localization Systems Security (Tracking of People and Goods), Mobile Commerce Security, Privacy Enhancing Technologies, Security Protocols (for Identification and Authentication, Confidentiality and Privacy, and Integrity), Ubiquitous Networks: Ad Hoc Networks Security, Delay-Tolerant Network Security, Domestic Network Security, Peer-to-Peer Networks Security, Security Issues in Mobile and Ubiquitous Networks, Security of GSM/GPRS/UMTS Systems, Sensor Networks Security, Vehicular Network Security, Wireless Communication Security: Bluetooth, NFC, WiFi, WiMAX, WiMedia, others

This Track will emphasize the design, implementation, management and applications of computer communications, networks and services. Topics of mostly theoretical nature are also welcome, provided there is clear practical potential in applying the results of such work.

Track B: Computer Science

Broadband wireless technologies: LTE, WiMAX, WiRAN, HSDPA, HSUPA, Resource allocation and interference management, Quality of service and scheduling methods, Capacity planning and dimensioning, Cross-layer design and Physical layer based issue, Interworking architecture and interoperability, Relay assisted and cooperative communications, Location and provisioning and mobility management, Call admission and flow/congestion control, Performance optimization, Channel capacity modeling and analysis, Middleware Issues: Event-based, publish/subscribe, and message-oriented middleware. Reconfigurable, adaptable, and reflective middleware approaches, Middleware solutions for reliability, fault tolerance, and quality-of-service, Scalability of middleware, Context-aware middleware, Autonomic and self-managing middleware, Evaluation techniques for middleware solutions, Formal methods and tools for designing, verifying, and evaluating, middleware, Software engineering techniques for middleware, Service oriented middleware, Agent-based middleware, Security middleware, Network Applications: Network-based automation, Cloud applications, Ubiquitous and pervasive applications, Collaborative applications, RFID and sensor network applications, Mobile applications, Smart home applications, Infrastructure monitoring and control applications, Remote health monitoring, GPS and location-based applications, Networked vehicles applications, Alert applications, Embedde Computer System, Advanced Control Systems, and Intelligent Control: Advanced control and measurement, computer and microprocessor-based control, signal processing, estimation and identification techniques, application specific IC's, nonlinear and adaptive control, optimal and robot control, intelligent control, evolutionary computing, and intelligent systems, instrumentation subject to critical conditions, automotive, marine and aero-space control and all other control applications, Intelligent Control System, Wiring/Wireless Sensor, Signal Control System. Sensors, Actuators and Systems Integration: Intelligent sensors and actuators, multisensor fusion, sensor array and multi-channel processing, micro/nano technology, microsensors and microactuators, instrumentation electronics, MEMS and system integration, wireless sensor, Network Sensor, Hybrid

Sensor, Distributed Sensor Networks. Signal and Image Processing: Digital signal processing theory, methods, DSP implementation, speech processing, image and multidimensional signal processing, Image analysis and processing, Image and Multimedia applications, Real-time multimedia signal processing. Computer vision, Emerging signal processing areas, Remote Sensing, Signal processing in education. Industrial Informatics: Industrial applications of neural networks, fuzzy algorithms, Neuro-Fuzzy application, bioInformatics, real-time computer control, real-time information systems, human-machine interfaces, CAD/CAM/CAT/CIM, virtual reality, industrial communications, flexible manufacturing systems, industrial automated process, Data Storage Management, Harddisk control, Supply Chain Management, Logistics applications, Power plant automation, Drives automation. Information Technology, Management of Information System: Management information systems, Information Management, Nursing information management, Information System, Information Technology and their application, Data retrieval, Data Base Management, Decision analysis methods, Information processing, Operations research, E-Business, E-Commerce, E-Government, Computer Business, Security and risk management, Medical imaging, Biotechnology, Bio-Medicine, Computer-based information systems in health care, Changing Access Patient Information. Healthcare Management Information Technology. to Communication/Computer Network, Transportation Application: On-board diagnostics, Active safety systems, Communication systems, Wireless technology, Communication application, Navigation and Guidance, Vision-based applications, Speech interface, Sensor fusion, Networking theory and technologies, Transportation information, Autonomous vehicle, Vehicle application of affective computing, Advance Computing technology and their application: Broadband and intelligent networks, Data Mining, Data fusion, Computational intelligence, Information and data security, Information indexing and retrieval, Information processing, Information systems and applications, Internet applications and performances, Knowledge based systems, Knowledge management, Software Engineering, Decision making, Mobile networks and services, Network management and services, Neural Network, Fuzzy logics, Neuro-Fuzzy, Expert approaches, Innovation Technology and Management: Innovation and product development, Emerging advances in business and its applications, Creativity in Internet management and retailing, B2B and B2C management, Electronic transceiver device for Retail Marketing Industries, Facilities planning and management, Innovative pervasive computing applications, Programming paradigms for pervasive systems, Software evolution and maintenance in pervasive systems, Middleware services and agent technologies, Adaptive, autonomic and context-aware computing, Mobile/Wireless computing systems and services in pervasive computing, Energy-efficient and green pervasive computing, Communication architectures for pervasive computing, Ad hoc networks for pervasive communications, Pervasive opportunistic communications and applications, Enabling technologies for pervasive systems (e.g., wireless BAN, PAN), Positioning and tracking technologies, Sensors and RFID in pervasive systems, Multimodal sensing and context for pervasive applications, Pervasive sensing, perception and semantic interpretation, Smart devices and intelligent environments. Trust, security and privacy issues in pervasive systems. User interfaces and interaction models, Virtual immersive communications, Wearable computers, Standards and interfaces for pervasive computing environments, Social and economic models for pervasive systems, Active and Programmable Networks, Ad Hoc & Sensor Network, Congestion and/or Flow Control, Content Distribution, Grid Networking, High-speed Network Architectures, Internet Services and Applications, Optical Networks, Mobile and Wireless Networks, Network Modeling and Simulation, Multicast, Multimedia Communications, Network Control and Management, Network Protocols, Network Performance, Network Measurement, Peer to Peer and Overlay Networks, Quality of Service and Quality of Experience, Ubiquitous Networks, Crosscutting Themes - Internet Technologies, Infrastructure, Services and Applications; Open Source Tools, Open Models and Architectures; Security, Privacy and Trust; Navigation Systems, Location Based Services; Social Networks and Online Communities; ICT Convergence, Digital Economy and Digital Divide, Neural Networks, Pattern Recognition, Computer Vision, Advanced Computing Architectures and New Programming Models, Visualization and Virtual Reality as Applied to Computational Science, Computer Architecture and Embedded Systems, Technology in Education, Theoretical Computer Science, Computing Ethics, Computing Practices & Applications

Authors are invited to submit papers through e-mail <u>ijcsiseditor@gmail.com</u>. Submissions must be original and should not have been published previously or be under consideration for publication while being evaluated by IJCSIS. Before submission authors should carefully read over the journal's Author Guidelines, which are located at http://sites.google.com/site/ijcsis/authors-notes.

© IJCSIS PUBLICATION 2010 ISSN 1947 5500